Network Theoretic Tools in the Analysis of Complex Diseases

Christopher Banerji

A thesis submitted to
University College London
for the degree of
Doctor of Philosophy

JUNE 2015

I, Christopher Banerji, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: .....

On the subjects of this thesis:

On Medicine: *We rationalise, we dissimulate, we pretend: we pretend that modern medicine is a rational science, all facts, no nonsense and just what it seems. But we have only to tap its glossy veneer for it to split wide open and reveal to us its roots and foundations, its old dark heart of metaphysics, mysticism, magic and myth. Medicine is the oldest of arts and the oldest of sciences: would one not expect it to spring forth from the deepest knowledge and feelings we have?* - Oliver Sacks, Awakenings, (1982).

On Mathematics: *Mathematics is not a book confined within a cover and bound between brazen clasps, whose contents it need only patience to ransack; it is not a mine, whose treasures may take long to reduce to possessions, but which fill only a limited number of veins and lodes; it is not a soil, whose fertility can be exhausted by the yield of successive harvests; it is not a continent or an ocean, whose area can be mapped out and its contour defined: it is limitless as the space which it finds too narrow for its aspirations; its possibilities are as infinite as the worlds which are forever crowding in and multiplying upon the astronomer's gaze; it is as incapable of being restricted within assigned boundaries or being reduced to definitions of permanent validity, as the consciousness, the life, which seems to slumber in each monad, in every atom of matter, in each leaf and bud and cell, and is forever ready to burst forth into new forms of vegetable and animal existence.* - James Joseph Sylvester, Speech made on Commemoration Day at John Hopkins University, (1877).

On Computing: *For it is unworthy of excellent men to lose hours like slaves in the labour of calculation, which could be safely relegated to anyone else if the machine were used.* - Gottfried Leibniz, *Machina arithmetica in qua non aditio tantum et subtractio sed et multiplicatio nullo, divisio vero paene nullo animi labore peragantur*, (1685).

Abstract

In this thesis we consider the application of network theoretic tools in the analysis of genome wide gene-expression data describing complex diseases, displaying defects in differentiation. After considering the literature, we motivate the construction of entropy based network rewiring methodologies, postulating that such an approach may provide a systems level correlate of the differentiation potential of a cellular sample, and may prove informative in the analysis of pathology. We construct, analytically investigate and validate three such network theoretic tools: Network Transfer Entropy, Signalling Entropy and Interactome Sparsification and Rewiring (InSpiRe). By considering over 1000 genome wide gene expression samples corresponding to healthy cells at different levels of differentiation, we demonstrate that signalling entropy is a strong correlate of cell potency confirming our initial postulate. The remainder of the thesis applies our network theoretic tools to two ends of the developmental pathology spectrum. Firstly we consider cancer, in which the power of cell differentiation is hijacked, to develop a malicious new tissue. Secondly, we consider muscular dystrophy, in which cell differentiation is inhibited, resulting in the poor development of muscle tissue. In the case of cancer we demonstrate that signalling entropy is a measure of tumour anaplasia and intra-tumour heterogeneity, which displays distinct values in different cancer subtypes. Moreover, we find signalling entropy to be a powerful prognostic indicator in epithelial cancer, outperforming conventional gene expression based assays. In the case of muscular dystrophy we focus on the most prevalent: facioscapulohumeral muscular dystrophy (FSHD). We demonstrate that muscle differentiation is perturbed in FSHD and that signalling entropy is elevated in myoblasts over-expressing the primary FSHD candidate gene *DUX4*. We subsequently utilise InSpiRe, performing a meta-analysis of FSHD muscle biopsy gene-expression data, uncovering a network of *DUX4* driven rewired interactions in the pathology, and a novel therapeutic target which we validate experimentally.

Overview

'In medicine - as in life - some problems are more complex than others and, science being the art of the soluble, it is only sensible to leave the apparently intractable aside, hoping perhaps that at some time in the future something will happen to open the doors to their resolution.' (James Le Fanu, The Rise and Fall of Modern Medicine [1]). Regardless of sensibilities, it is clear that patients with complex pathology cannot wait for some undefined future time. Moreover, since these words were written the age of multidisciplinary science has caused 'complexity' to take on an altogether different meaning. The vision of a complex biomedical problem as intractable has been usurped, in its place we have a perturbation to a large and sophisticated system. Understanding such a perturbation requires a twofold investigation. First: the generation of high throughput data describing the detailed interplay of interacting components, and second: the development and application of mathematical models intended to convert such data into a mechanistic understanding of the pathology. Given the abundance of publicly available high-throughput data, in this thesis we will dominantly focus on the development and application of mathematical models for the understanding of complex disease. Pathology-wise our key interests will be cancer and facioscapulohumeral muscular dystrophy (FSHD), for the latter pathology we will validate some of our computational findings experimentally.

Following introduction, we, in the second chapter, describe the development of three network theoretic tools for investigating genome wide gene expression data. The first, *Network Transfer Entropy* (NTE) proposes a model of signal transduction, which can be represented as a stochastic matrix, describing the interaction probability of network vertices. The model is theoretically examined from a metric space perspective, and a measure is constructed for comparing dynamics on two weighted networks. The measure is then refined to compute the information transferred between any two vertices in a weighted biochemical network and its validity is demonstrated by application to a phosphorylation network. The second methodology, *signalling entropy*, further considers the signal transduction model applied for the development of NTE, exploring the entropy rate of the stochastic matrix as a global measure of network robustness. It is analytically demonstrated that signalling entropy is a population level correlate of intra-sample heterogeneity and that it is driven by the activity of highly connected vertices. The final methodology *Interactome Sparsification and Rewiring* (InSpiRe), considers the signal transduction model again on a local scale, to construct an algorithm for identifying network rewiring between two biological phenotypes.

After construction, the second are third network theoretic tools are applied in a number of settings. As development is frequently altered in complex pathology, in the third chap-

ter, we first motivate and demonstrate that *signalling entropy* is a robust and powerful measure of the differentiation potential of a single cellular sample. Through the consideration of over 1000 genome wide gene expression assays we reveal that signalling entropy systematically decreases during differentiation. Moreover, we find that the local entropy measure utilised in InSpiRe is capable of identifying genes critical to the differentiation process. We also consider signalling entropy in the context of pathological development, confirming findings that our measure is elevated in cancerous compared to healthy tissue, and demonstrate that it is elevated in cancer stem cells compared to the tumour bulk.
In the fourth chapter we investigate the relationship between signalling entropy and cancer in more detail. We demonstrate that our measure correlates with many quantifiers of tumour stemness across multiple malignancies. Following this, we reveal that signalling entropy is strongly prognostic in both breast and lung cancer, by considering over 5000 bulk tumour samples. We also reveal a number of intriguing associations between signalling entropy and cancer clinical variables, including that signalling entropy is elevated in lung cancer patients with a history of smoking, and in luminal B beast cancer patients despite the low expression of stem cell genes. These results suggest a differential role for intra-tumour heterogeneity and stemness in certain cancer subtypes. We further demonstrate that the prognostic power of signalling entropy is superior to other gene expression based indicators, and that it is driven by genes involved in cancer stem cells and treatment resistance.
Finally, in the fifth chapter, we consider FSHD as a pathology of cellular differentiation, independent to cancer. We investigate the primary FSHD candidate gene *DUX4* and reveal that signalling entropy is elevated in mouse myoblasts over-expressing this gene. This result is indicative of inhibited differentiation driving this pathology. We subsequently consider a human myoblast model of FSHD and demonstrate that *DUX4* is over expressed at low levels, leading to an inhibited muscle differentiation programme. To further investigate the molecular mechanisms driving FSHD we utilise the InSpiRe algorithm. By performing a meta-analysis of publicly available muscle biopsy gene expression data sets, we reveal a detailed network of rewired interactions in FSHD. We show that the expression of genes in this network are modified by *DUX4* expression confirming this candidate as driver of FSHD pathomechanisms. Interrogation of this network by betweenness centrality reveals β-catenin signalling as critical to pathomechanisms. We subsequently validate experimentally that β-catenin is activated by *DUX4* expression.

Facilities

1. Centre of Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London WC1E 6BT, UK.

2. Department of Computer Science, University College London, London WC1E 6BT, UK.

3. Statistical Cancer Genomics, Paul O'Gorman Building, UCL Cancer Institute, University College London, London WC1E 6BT, UK.

4. Randall Division of Cell and Molecular Biophysics, New Hunt's House, King's College London, Guy's Campus, London SE1 1UL, UK.

Publications

Much of the work described in this thesis has been published in peer-reviewed academic journals, with some work still in preparation. The adapted contents of the following publications feature in this thesis (* denotes corresponding author):

1. C. R. S. Banerji*, S. Severini and A. E. Teschendorff, Network transfer entropy and metric space for causality inference, *Phys. Rev. E*, **87**:052814 (2013), (pdf:arXiv:1303.0231).

2. C. R. S. Banerji, D. Miranda-Saavedra, S. Severini, M. Widschwendter, T. Enver, J. X. Zhou, A. E. Teschendorff*, Cellular network entropy as the energy potential in Waddington's differentiation landscape, *Scientific Reports*, **3**:3039, (2013).

3. C. R. S. Banerji*, S. Severini, C. Caldas and A. E. Teschendorff*, Intra-Tumour Signalling Entropy Determines Clinical Outcome in Breast and Lung Cancer. *PLoS Comput Biol*, **11**(3):e1004115, (2015).

4. A. E. Teschendorff*, C. R. S. Banerji, S. Severini, Reimer Kuehn and Peter Sollich, Increased signaling entropy in cancer requires the scale-free property of protein interaction network, *Scientific Reports*, in press (2015), (pdf: arxiv:1504.00120).

5. C. R. S. Banerji*, P. Knopp, L. Moyle, S. Severini, R. Orrell, A. E. Teschendorff, P. S. Zammit*, β-catenin is central to *DUX4* driven network rewiring in facioscapulo-humeral muscular dystrophy, *Journal of The Royal Society Interface*, **12**(102):20140797, (2015).

6. P. Knopp, C. R.S. Banerji, L. Moyle, J. Davies and P. S. Zammit*, *DUX4* inhibits satellite cell myogenesis and maintains a stem cell phenotype, *in preperation*

The author also contributed to the following publications which feature less prominently in the thesis:

7. C. R. S. Banerji, T. Mansour, and S. Severini*, A notion of graph likelihood and an infinite monkey theorem, *Journal of Physics A: Mathematical and Theoretical*, **47**(3),035101, (2014) (pdf: arXiv:1304.3600).

8. M.-L. Lin, H. Patel, J. Remenyi, C. R. S. Banerji, M. Pariyasamy, C.-F. Lai, S. Ottaviani, P. R. Quinlan, C. A. Purdie, L. B. Jordan, A. M. Thompson, R. C. Coombes, F. V. Fuller-Pace, A. E. Teschendorff, L. Buluwela, S. Ali*, Gene Expression Profiling of Nuclear Receptors as Diagnostic and Therapeutic Targets in Breast Cancer, *Oncotarget*, in press,(2015).

Acknowledgements

I would like to thank my supervisors Simone Severini, Andrew Teschendorff and Peter Zammit, for their incredible support, expert guidance and huge enthusiasm throughout the work on this thesis, as well as for the great intellectual freedom they permitted me in my research.

In addition, I would like to thank my collaborators and co-authors: Richard Orrell, Louise Moyle, Paul Knopp, Nico Figac, Congshan Sun, Pedro Quiroga, Robert Knight, James West, Toufik Mansour, Carlos Caldas, Diego Miranda-Saavedra, Martin Widschwendter, Tariq Enver, Joseph Zhou, Simak Ali, Meng-Lay Lin, Reimer Kuehn and Peter Sollich. Their diverse expertise, ranging from the clinical and experimental to the physical and mathematical sciences has been invaluable, and contributed greatly to the work presented in this thesis.

I must also thank my funders the Engineering and Physical Sciences Research Council (EPSRC), the British Heart Foundation (BHF) and the FSH Society, without the support of whom this work would not have have been possible.

Lastly I must thank family and friends, in particular, my parents Kathy and Ranjan, my sister Christina, and my girlfriend Catherine, for their interest, patience and encouragement and for dragging me from the ivory tower before cabin fever set in.

Contents

List of Figures

1 Introduction

Our field of study is multifaceted, interdisciplinary and fast paced, thus a review of all topics relevant to our work will almost surely be incomplete. However, I hope that in this introduction we can provide sufficient background to motivate our research. Following this introduction, the body of the thesis will consist of 4 chapters:

2. Entropic Network Theoretic Tools: Concept and Theory.
3. Signalling Entropy as the energy potential of Waddington's Landscape.
4. Signalling Entropy correlates with Clinical Outcome in Epithelial Cancer.
5. Network Rewiring in Facioscapulohumeral Muscular Dystrophy.

In the second chapter we describe a general model for dynamics on a weighted biological network, which can be derived from genome wide gene expression data. From this model we derive three network theoretic tools for investigating and understanding genes and pathways perturbed in complex pathology: *Network Transfer Entropy* (NTE), *Signalling Entropy* and *Interactome Sparsification and Rewiring* (InSpiRe). We investigate these tools analytically and demonstrate some general properties.
In the third chapter we consider cellular differentiation, a highly sophisticated process critical to multicellular organism development, which is perturbed in many complex pathologies. We demonstrate that one of our network theoretic tools, signalling entropy, is a powerful measure of the progression of differentiation and motivate the use of the measure in the understanding of diseases of development. In an oncogenic context we demonstrate signalling entropy is elevated in cancer stem cells as compared to the tumour bulk.
In the fourth chapter we investigate signalling entropy in more detail in the context of epithelial cancer. We reveal our measure to be correlated with tumour stemness and other clinical variables across multiple malignancies and strongly prognostic in both breast and lung cancer.
Finally, in the fifth chapter we investigate network rewiring in *facoiscapulohumeral muscular dystrophy* (FSHD), a complex disease of development. We first reveal the signalling entropy is raised in samples over-expressing the primary FSHD candidate gene *DUX4*. Subsequently, we experimentally demonstrate that *DUX4* is over-expressed in human muscle precursor cells from FSHD patients, and that such cells exhibit clear defects in muscle regeneration, such as inefficient alignment and fusion, leading to atrophic myotubes. To investigate in more detail which genes and pathways are driving FSHD pathomechanisms, we employ InSpiRe, our more local network theoretic algorithm, in a

meta-analysis of FSHD muscle biopsy gene expression data sets. We uncover a network of rewired protein interactions in FSHD, which can be attributed to *DUX4* expression. Betweenness centrality indicates that β-catenin is the most critical bottleneck to FSHD signalling. We subsequently validate that over-expression of *DUX4* induces dysregulation of a number of β-catenin downstream targets.
Given these topics of interest, we will focus our introduction on the following areas:

1. The development of network theoretic tools.
2. Mathematical approaches to understanding cell differentiation
3. The pathophysiology of breast cancer.
4. The pathophysiology of FSHD.

At the end of each introductory review we will consider the lessons learned from the literature and explain the relevance of the information to guiding our research.

1.1 The development of network theoretic tools

"Genes do not act in isolation" [2]. This is a simply stated fact but the ramifications are profound; for the complex and detailed interplay between intracellular components forms the basis of the most sophisticated and enigmatic biological processes, and the perfect hiding place for elusive mechanisms of incurable disease.
Unmasking these elaborate molecular mechanisms in healthy and diseased biological processes has thus become a rich sub-discipline of systems biology. The complexity of this field is embodied by its demand for strong inter-disciplinary, with the deepest insights requiring a powerful combination of well-designed, high throughput experimental biology, with the computational and mathematical sciences to analyse and interpret the resulting data.
Large-scale ventures in the study of the molecular mechanisms of the cell often focus on the representation and analysis of biological pathways as large graphs, referred to as networks, in which vertices represent interacting biological components connected by fixed edges [3]. Many algorithms have been designed and refined for the purpose of reverse engineering these networks from genome wide gene expression data [4]. Early approaches focused on utilising hierarchical clustering or co-expression to postulate gene interdependencies. These were largely succeeded by more sophisticated methodologies, drawing on the concepts of information theory [5], Bayesian networks [6] and dynamical systems [7]. Debate still persists on the superiority of these more recent approaches, and it seems

likely that the optimal choice of network inference algorithm is dependent on a number of variables [4]. In addition, many large, online, manually curated data bases exist, detailing experimentally verified interactions for the empirical construction of networks [8, 9, 10, 11, 12] and many investigators prefer to work with these more high confidence interaction maps, to overcome the insufficiencies of inference algorithms.

Subsequent analysis of these networks relied heavily on the classical field of graph theory [13]. Topological measures of graph centrality became a popular approach to the interrogation of structural properties of interacting biological systems, frequently providing insight into genes which are well placed to drive malicious phenotypes [3]. More recently some exciting approaches in the analysis of network structure have arisen from the consideration of geometric graphs and the statistical embedding of a biological network into a metric space [14].

Further insights on the nature of genetic interactions have been achieved by the integration of genome wide expression data with biological networks to generate weighted graphs [15, 16, 17]. Methodologies to analyse these richer structures often draw insight from stochastic analysis, most notably from random walk theory, to construct more informative network centrality measures and to interrogate dysregulated dynamics in disease [15, 18, 19].

Recently, the study of large scale molecular interactions has undergone a paradigm shift, following the discovery that the structure of the complex cellular machinery can be entirely re-organised in response to environmental perturbation, disease or even naturally during development [20, 21, 22]. This insight, dubbed network rewiring, has led to the construction of a new set of mathematical and computational tools which have been designed to investigate experimental data from this more informative viewpoint.

These more recent methodologies have generally focused on the comparison of two or more weighted networks, describing different phenotypes or conditions. Among the most successful approaches are random walk based algorithms, which have led to significant advances in our understanding of sophisticated biological processes, revealing novel mechanisms for complex diseases and elucidating drug responses [16, 17, 23].

In what follows we will review popular methodologies employed for the analysis of health and disease from the perspective of complex networks. We will progress systematically, first considering how one constructs a network of biological interactions by experimental approaches and via statistical methodologies for gene network inference from expression data. We then progress to explain the analysis of such networks from a single phenotype perspective. Subsequently, we consider differential networks and chart the rise of the powerful concept of network rewiring, describing the recent methodologies developed around

this notion and the new insights gained. We close with some perspectives, explaining the relevance of the reviewed literature to our research goals and examining the flavour of methodology that given insights from our findings would be useful to develop.

1.1.1 Uncovering the Structure of Biological Networks

In order to understand and interrogate the structural and dynamical properties of biological networks, one must first have a reasonable knowledge of at least a representative subset of bio-molecular interactions. Many biochemical assays exist to identify such interactions with a high sensitivity and specificity. The most utilised of these are the immunological methods such as *co-immunoprecipitation* (Co-IP) and *chromatin immunoprecipitation* (ChIP). Co-IP utilises antibodies against specific proteins to investigate putative interactions, and can even be used to identify unknown interaction partners for a specified protein [24], whilst ChIP precipitates DNA fragments to which a protein of interest binds, to identify genetic regulatory interactions. Whilst accurate, these methodologies are reasonably small scale and can only investigate the interaction partners of few proteins at a time, and under specific laboratory conditions. More high-throughput experimental techniques for interaction identification exist, such as *Yeast 2 Hybrid* (Y2H), which utilises a coupled reporter gene system in yeast to identify gene pair interactions [25]. Whilst useful, Y2H is renowned for false positives, likely caused by the natural separation of two proteins in the organism of interest (by organelles etc.) which is not recapitulated in the experimental setting. Though this is an issue, annotations describing intra-cellular localisations for many proteins exist [26] and can be used to refine Y2H identified interactions and remove false positives. Collectively, the results of these sophisticated molecular biological techniques have been compiled into a number of evidence based interaction data bases, such as Pathway Commons, BioGrid, IntAct and KEGG [8, 9, 10, 11, 12], which serve as high confidence resources for network data.

In addition to the use of targeted assays, the large scale identification of molecular interactions has been achieved by reverse engineering genome wide gene expression data into genetic networks, using a multitude of different mathematical approaches [4, 5, 6, 27, 28]. Underlying all these algorithms is the principle that if a gene x is perturbed, resulting in the modification of the expression of another gene y, then x and y must either directly or indirectly (through some mediators) interact. Thus if one is given an expression data set comprising global gene expression profiles for an organism, under a large number of known perturbations, theoretically one will be able to uncover causal interdependencies between genes, with the caveat that certain relations will be indirect. In reality such a perfect data set does not exist [27]. Rather gene expression data generally consists of either a

number, M, of observations of the expression of many, n, genes either in steady state, or ordered as a time course. Perturbations between the observations are generally unknown and may derive from genetic variation between samples in the case of the steady state observations [4]. Consequentially, causal interdependencies between genes are difficult to infer as interactions cannot generally be traced from a causal perturbation. However, this does not mean that network inference is impossible [29], rather interactions must be identified through measures designed to infer dependencies between the expression profiles of gene pairs from the M samples [4, 27].

A popular methodology for inferring genetic networks from expression data is *Algorithm for the Reconstruction of Accurate Cellular Networks* (ARACNe), which employs an information-theoretic approach [5]. ARACNe treats the expression of each gene as a random variable and utilises mutual information, between data estimated gene expression probability density functions, to assess interdependencies between gene pairs. The mutual information between two random variables X and Y can be defined in terms of entropies

$$I(X,Y) := S(X) + S(Y) - S(X,Y),$$

where $S(X) := -\sum_x P_X(x) \log P_X(x)$ is the entropy of X,
$S(X,Y) = -\sum_{x,y} P_{X,Y}(x,y) \log P_{X,Y}(x,y)$ is the joint entropy of X and Y and $P_X(x)$ and $P_{X,Y}(X,Y)$ are the probability mass function (pmf) of X and the joint pmf of X and Y respectively. $I(X,Y)$ is a symmetric, model free measure of the interdependency of X and Y, in particular if X and Y are independent variables, i.e., $P_{X,Y}(x,y) = P_X(x)P_Y(y)$, then $I(X,Y) = 0$. ARACNe uses a Gaussian Kernel estimator [30] to approximate the joint probability distribution of the expression of all gene pairs described by the expression data, from which the univariate distributions are computable as the marginals. The mutual information between every gene pair is thus computed and a randomisation procedure is utilised on the expression data to identify a mutual information threshold above which gene interdependency is statistically significant. ARACNe also employs the *Data Processing Inequality* (DPI) to eliminate potentially non-causal interactions. The DPI states that for three genes in a network X, Y and Z, if the edges (X,Y) and (Y,Z) exist in the true network but (X,Z) does not, then it follows that

$$I(X,Z) \leq \min[I(X,Y); I(Y,Z)].$$

It should be noted that this is a necessary but not sufficient condition and that by employing the DPI, ARACNe may be deleting a number of real interactions [4].

ARACNe, has been shown to be successful in inference of networks from in silico data [4, 5], most notably in the case of small data sets of steady state expression profiles, where

it out performs many other popular methodologies. In order for the mutual information to be well defined, however, ARACNe requires the input expression data assays to be mutually independent measurements; consequentially ARACNe has demonstrated poor performance on short spaced, time series data where this assumption fails [4]. Moreover, as mutual information is a symmetric measure, ARACNe can only ever produce undirected networks. Despite these limitations the methodology has provided strong insights into biological networks. Notably, in an analysis of expression data from 336 perturbations of human B cells, ARACNe identified a large number of targets of the proto-oncogene *MYC*, including many which had not been previously documented [31]. By performing ChIP, the authors of this study validated many of these putative targets, confirming the ability of the methodology to predict true biological interactions. Moreover, many of the downstream targets of *MYC* were found to be highly connected in the ARACNe inferred network, suggesting a hierarchical nature to genetic regulation [28].

Another class of reverse engineering algorithms include methodologies such as *Bayesian Network Inference with Java Objects* (BANJO) [6], which utilise Bayesian networks to infer probabilistic relationships between genes. In a Bayesian network interdependencies between a set of random variables $X_1, ..., X_n$ are assumed to be in the form of a directed acyclic graph $\mathcal{G} = (V, E)$, where $V := \{X_1, ..., X_n\}$ is a set of vertices and $E \subset \{(i, j) : i, j \in V\}$ is a set of directed edges. A Markovian assumption is also employed, stating that a variable X_k is dependent only on its immediate inputs in $\mathcal{G}$, also known as the parental variables of X_k. This assumption, via Bayes theorem, can be represented by a probability distribution $P(X_1..., X_n)$ satisfying

$$P(X_1..., X_n) = \prod_{i=1}^{n} P(X_i | \{X_j : (j, i) \in E\}.$$

In the reverse engineering of gene networks, each variable X_k can be considered the expression of a gene, and the M repeated measurements of gene expression at steady state represent a sample of these variables, which we will denote by Y. The task of Bayesian network algorithms is thus to infer the optimal $\mathcal{G}$ for the data set Y, generally using a maximum likelihood approach. This amounts to finding $\mathcal{G}$ which maximises $P(\mathcal{G}|Y)$, which can be represented via Bayes rule as

$$P(\mathcal{G}|Y) = \frac{P(Y|\mathcal{G})P(\mathcal{G})}{P(Y)},$$

where $P(\mathcal{G})$ may encode some prior on the expected graphical structure (possibly guided by literature curated networks [32]). Finding the precise $\mathcal{G}$ which maximises $P(\mathcal{G}|Y)$, will likely lead to over-fitting due to the large parameter space represented by $\mathcal{G}$ however, thus Bayesian Scoring metrics, with complexity penalizations are generally employed. Due to

the huge state space of possible networks $\mathcal{G}$, it is impractical to test every possibility to identify the structure that maximises the score. Consequentially, a variety of search algorithms have been utilised to sample the parameter space in a manner which permits an approximation of the optimal network structure. Popular examples of these algorithms include greedy search, simulated annealing and genetic algorithms. It was demonstrated that in the BANJO algorithm setting, all three search algorithms gave similar results on simulated data, however the greedy search provided the most rapid convergence [6].

The Bayesian network algorithm BANJO was directly compared to ARACNe in terms of sensitivity and specificity of networks inferred from simulated data [4]. This revealed that BANJO is very powerful for the inference of smaller networks (< 1000 vertices) when there is a large amount of data available (> 10 samples). However, the complexity of BANJO prevented it from converging on larger networks, and in the cases when datasets were small, ARACNe was a clear winner, with BANJO performing no better than random. Despite this limitation, Bayesian network approaches to the reverse engineering of genetic interactions have provided many important results [32], particularly through the use of the prior $P(\mathcal{G})$ to encode additional information to the expression data, guiding the construction of more accurate networks. For example, Jensen *et al.* [33] utilised ChIP binding data to inform the prior and improve network inference in a reverse engineering of a gene regulatory network from expression data in yeast.

In addition to Baysean networks and information theoretic approaches, a third class of reverse engineering algorithms takes its origins from the field of dynamical systems. In these algorithms, the expression and interactions between a set of genes $X_1, ..., X_n$ are modelled by a deterministic set of ordinary differential equations (ODEs)

$$\frac{dX_k}{dt} = f(X_1, ..., X_n, \theta),$$

where f is a continuous function and θ is a vector of parameters which specifies the network. The expression data is then utilised to estimate θ, either by setting $\frac{dX_k}{dt} = 0$ in the case of steady state data, or estimating the time derivative in the case of time course data. The choice of f is clearly critical to this approach, with non-linearity providing more realism at the cost of higher computational complexity. Despite this trade off, it appears that a linear model may be sufficiently powerful for reliable inference and many popular algorithms such as *Network Inference by Reverse-engineering* (NIR) [7], utilise such a linear model and regression techniques to successfully engineer networks from expression data. The deterministic nature of these *ordinary differential equation* (ODE) approaches, however, mean that they are better suited to data sets where noise is low [4]. Thus a combination of experimental and computational techniques can be employed to

identify large scale genetic interactions within the cell, and map out molecular networks for detailed interrogation. There is a vast array of algorithms available for the reverse engineering of interaction networks from expression data, each with their own strengths and weaknesses that depend critically on the data set being employed for inference. In this work we will aim to develop general methodologies that are useful regardless of data set structures. Hence we will not restrict ourselves to a given network inference algorithm which may provide differing results on independent data sets describing the same phenotype. Instead we will utilise the experimentally compiled networks, described by the integrated results of the many online manually curated interaction data-bases. As explored above the reliance of these data bases on Y2H experiments may lead to false positives due to poor localisation restrictions, hence we will utilise protein intra-cellular localisation annotations to refine our network. It should be noted that an advantage of gene network inference is that its derivation from data provides a sample specific context to the interaction patterns. We will thus only consider our experimentally derived networks as weighted graphs, where edge weights are derived from gene expression data and provide a sample specific context to the known interactions. In this way we create sample specificity without introducing/removing interactions in a manner which may over fit the network to the data set.

Once the network structure is decided upon, the next stage of analysis requires the development and implementation of methodologies to interrogate the global and local nature of biological networks, with the aim of identifying critical mediators of healthy and pathological processes. Early approaches to this investigation began via the evaluation of the topological properties of biological networks, describing either general biological interactions or interactions attributable to a single biological phenotype.

1.1.2 Single Phenotype Networks

Once a biological network was elucidated describing a single phenotype, either via experimental or algorithmic means, the challenge became to understand how the structure of the network could inform us about the system it described. Early characterisation of inferred and experimentally verified networks focused on centrality measures of the static network topology and famously suggested a scale free structure [31, 34, 35]. In scale free networks the degree distribution typically follows an inverse power law, resulting in a network in which highly connected vertices or hubs are rare and poorly connected vertices predominate (Fig 1.1). Scale free networks became highly studied and were found to be robust, in terms of connectivity, to the deletion of random vertices, but particularly sus-

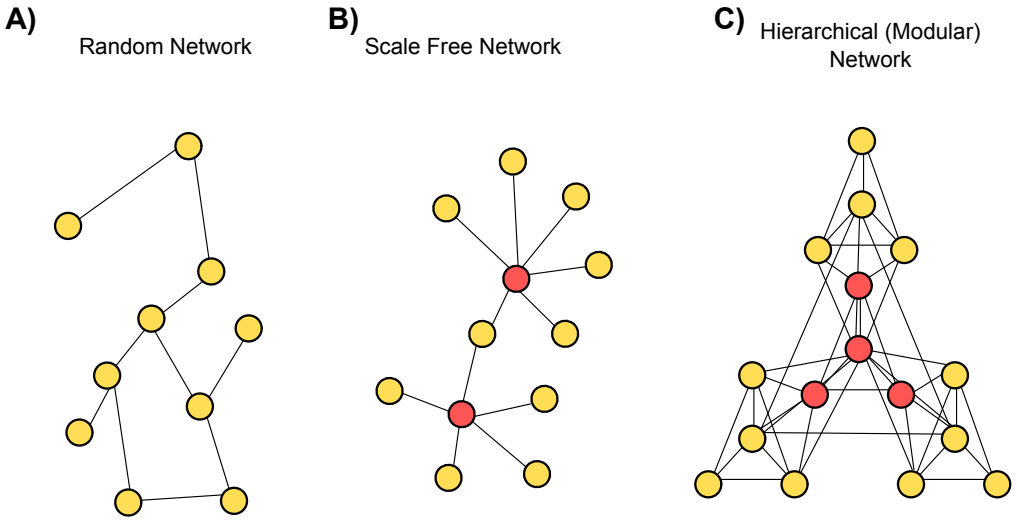

Figure 1.1: **Network topologies.** (A) Random networks have normally distributed degree distributions. (B) Scale free networks have power-law degree distributions, resulting in a low number of high degree vertices. (C) Hierarchical networks are scale-free and yield a modular structure, with high degree vertices displaying higher clustering coefficients (forming more triangles via their edges). Biological networks have been posited to display concordance with a hierarchical network structure.

ceptible to the targeted deletion of highly connected vertices [36]. The biological relevance of this theory was subsequently demonstrated in yeast, where it was shown that knock outs of genes corresponding to hubs in the protein interaction network, were far more likely to convey lethality than poorly connected genes [37]. However, our understanding of biological interactions is constantly evolving and, for the present at least, we only ever have knowledge of a subset of all true biological interactions. This is concerning as it was recently demonstrated that subsets of scale free networks are not likely to be scale free [38]. Moreover, more current studies of static networks compiled from publicly available data bases have revealed a troubling inconsistency with previously established scale free structure [39, 40].

Further insights into biological network structure came from the identification of modularity, deriving from the initial observation, in yeast, that genes with similar functional annotations were more likely to cluster together in protein-protein interaction networks [41] (Fig 1.1). This revealed modularity of biological networks has motivated more local analysis of topology, and led to methodologies permitting the decomposition of networks

into pathways and subsystems for detailed functional analysis [42]. Modularity has also led to the use of a different class of centrality measures, aimed at identifying inter-modular hubs, which may not be highly connected but are clear bottlenecks for information flow between discrete modules in the network. Among these measures is the well-studied betweenness centrality, which quantifies the number of shortest paths between any two vertices in the network which pass through a given vertex. It was demonstrated that this measure is computable through the consideration of the stationary distribution of a walker moving randomly through a network [43]. These inter-modular hubs were found to be mutated in breast cancer far more frequently than intra-modular hubs [44], demonstrating the importance of communication between functional network subsets for the maintenance of healthy homoeostasis.

The analysis of single phenotype biological network structure has also diverged from the consideration of local centrality measures to understand more global properties. It was recently observed that complex networks display self-similarity, particularly with respect to degree-threshold re-normalisation (considering the ensemble of complex networks generated by sequentially thresholding the degree of vertices) [45]. The main topological characteristics of the networks in the ensemble generated by this procedure (*e.g.* degree distribution, clustering coefficient) are found to be self-similar, a property not observed in randomised networks. It was subsequently demonstrated that if one considers a graph generated by uniformly distributing vertices in a metric space, and then connecting vertices with a probability inversely proportional to their separation, then one can generate from this assumption networks which demonstrate such self-similarity. By employing a maximum likelihood approach, Serrano *et al.* [14] embedded the topologies of metabolic networks into a hyperbolic geometry and found that reactants with functional similarity tended to be more proximal in this underlying metric space, implying a hidden modularity is encoded in the structure of these networks.

The study of single phenotype networks advanced further from the consideration of weighted network topologies. This permitted the use of methodologies based on the notion of random walks on weighted graphs, which provided insight into disease pathomechanisms. Among these methodologies is NetRank, a modification of the Google PageRank algorithm, which was able to identify novel, robust, network based biomarkers for survival time in various cancers [18, 19]. The methodology integrates gene expression data, survival data and interaction data describing a single cancer phenotype to generate a weighted network. Briefly, each vertex in a biological network is matched to a gene in the expression data and then weighted with the correlation between gene expression and survival time. The algorithm then considers the motion of a random walker on the

weighted biological network, at each time point the walker can move to a neighbouring vertex with probability d, with the choice of neighbour uniform, or to a random vertex in the network with probability $(1-d)$ with the choice of random vertex proportional to the vertex weight (correlation of expression with survival time). In this way the walker is biased towards the neighbourhood of genes which are over-expressed in patients with a better clinical outcome and away from the neighbourhood of genes which are under-expressed. This process is iterated until convergence to obtain a ranking for the vertices in terms of visitation probability of the random walker. To compute this final ranking an iterative approach is adopted by noticing that the ranking r_j^k of the j^{th} vertex in the network after the k^{th} evolution of the random walker satisfies

$$r_j^k = (1-d)c_j + d\sum_{i=1}^{V} \frac{a_{ij} r_i^{k-1}}{deg(i)},$$

where a_{ij} is the adjacency matrix of the network, c_j is the correlation of the expression of the gene corresponding to vertex j with survival time, V is the number of vertices in the network and $deg(i)$ is the degree of vertex i in the network. Consequentially, at convergence the ranking satisfies

$$(I - dA^T D^{-1})\vec{r} = (1-d)\vec{c},$$

where I is the identity matrix, A is the adjacency matrix of the network, D is a diagonal matrix of vertex degrees and $\vec{r}$ and $\vec{c}$ are the vectors of rankings and correlations of expression with survival time respectively. The top ranked features are then utilised to train a classifier for predicting survival time in various cancers. Estimating the optimal d for NetRank has been achieved by Monte Carlo cross validation, in which the classifier training data set is split randomly into a test and training set and various values of d are utilised to train and test a classifier, with the optimal value of d then chosen as that which leads to the classification with the highest accuracy.

Thus the analysis of single phenotype biological networks, through the consideration of topology and random walks has led to many insights on how form may influence function in interacting biological systems. However, these methodologies ignore a potential dynamism between different biological phenotypes; a network rewiring which may provide deeper insights into the differences across biological phenotypes and how to ameliorate them.

1.1.3 Network Rewiring: a shift in methodology

Whilst the analysis of single phenotype biological networks has proved insightful in the identification of critical disease genes, recent studies have revealed that the true topology of biological networks is far from static; rather it is constantly rewiring in response to environmental and genetic perturbations [46]. Such a revelation suggests that networks underlying biological phenotypes must be considered in a case-control setting, and that the rewired differences between networks describing phenotypes are of more critical importance than the phenotype specific network topologies. The concept of network rewiring in response to external stimulus was most convincingly demonstrated in the work of Bandyopadhyay *et al.* [20]. The authors of this study used a new technique dubbed differential epistatic mini-array profiling, to identify the remodelling of all pairwise genetic interactions, between 418 genes in yeast, following exposure to the DNA damaging agent methanesulfonate (MMS). Briefly, the methodology involved generating a large number of pairwise gene knockouts (or hypomorphic alleles for essential genes), one for each gene pair assessed. Each pairwise interaction was then scored, in each condition (exposure or non-exposure to MMS), with a measure of cell viability. The authors thus generated two weighted networks, one describing genetic, cell viability modifying interactions in standard laboratory conditions and the other describing interactions under exposure to MMS. The two networks showed large differences, though intriguingly the MMS treated network was not enriched for interactions between DNA damage response genes. However, when the untreated network was subtracted from the MMS treated network, significant enrichment of DNA damage response was revealed. This discovery emphasised that it is in the differential interactions between two conditions that the biological response can be found, rather than in the static network of the responding organism (Fig 1.2). This finding has contributed to a rapid growth in the study of differential networks, and to the prominence of the concept of network rewiring [47].

Understanding of genetic disease has benefited from the consideration of network rewiring, where a causal genetic aberration leads to a restructuring of intra-cellular interactions from the healthy state. Many early studies considered a gene mutation as the deletion of a vertex in a static network with no further rewiring. Such an approach is limited, however, as it cannot explain how multiple diverse diseases may arise from mutations in the same gene. It was subsequently demonstrated that many mutations corresponding to Mendelian diseases, may in fact be better represented by deletion or remodelling of network interactions [22]. In this model mutations in a single gene could naturally lead to many diverse diseases, through subtle modification of interaction preferences.
A focus on the concept of network rewiring has led to advances in the study of patholo-

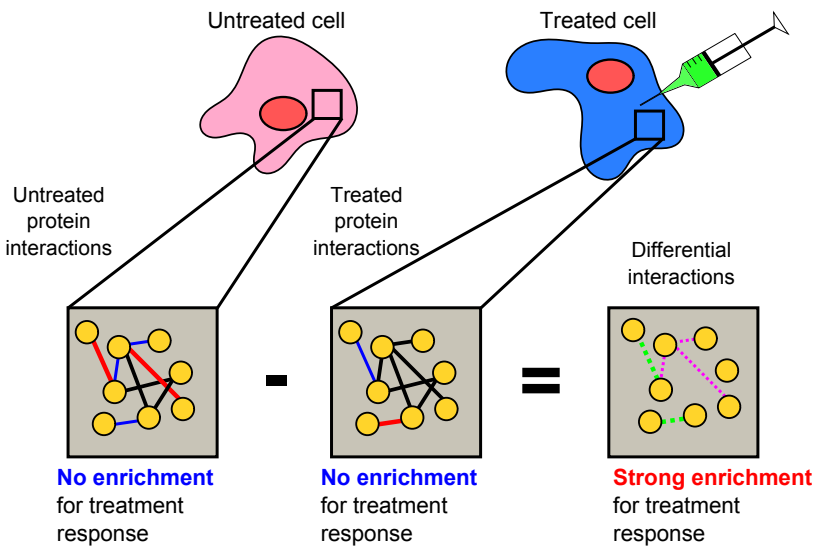

Figure 1.2: **Network Rewiring.** Protein interaction networks dynamically rewire in response to treatment with DNA damaging agents. The static protein interactions within treated and untreated cells are not enriched for treatment response, however the differential interactions are strongly enriched.

gies from the consideration of differential co-expression networks [27]. In such studies gene expression data profiling healthy and diseased tissue is used to reverse engineer a weighted network for each phenotype utilising ARACNe or BANJO. Topological metrics or clustering techniques are then utilised to identify genes and interactions which are substantially remodelled in disease. This strategy has been successfully utilised to identify gene sets activated and repressed during mouse mammary tumour gland formation [48]. In addition, differential co-expression networks have been able to identify oncogenic transcriptional regulators, to which conventional differential expression techniques are blind [49].

Random walk theory has also been employed in the development of differential network methodologies, such as NetWalk [16], which investigate network dynamics. This methodology is similar to the single phenotype NetRank but investigates the behaviour of the random walker in a differential setting. NetWalk first integrates gene expression data with a compiled network of all known biological interactions referred to as the Knowledgebase (with adjacency matrix $(a_{ij})_{ij=1}^{V}$, where V denotes the number of vertices), by utilising a general vertex weighting. Every vertex i in the network is assigned a value w_i, which may correspond to any number of data derived quantities such as the differen-

tial expression of the gene corresponding to vertex i between two phenotypes of interest. As in NetRank, the vertex weightings are then utilised to construct a stochastic matrix $(p_{ij})_{ij=1}^{V}$, describing the transition probabilities of a random walker biased by the vertex weights $(w_i)_{i=1}^{V}$:

$$p_{ij} = (1-q)\frac{a_{ij}w_j}{\sum_{k\in\mathcal{N}_i} w_k} + \frac{qw_j}{\sum_{k=1}^{V} w_k},$$

where $\mathcal{N}_i$ is the neighbourhood of vertex i in the network and $q \in (0,1)$. By the Perron Frobenius theorem, the left eigenvector of this transition probability matrix corresponds to the stationary distribution, $(r_i)_{i=1}^{V}$, of the walker. In addition to computing $(r_i)_{i=1}^{V}$, NetWalk also considers an edge flux, e_{ij}, which is the probability of the walker (initiated from its stationary distribution) traversing an edge (i,j) in the network, simply evaluated as:

$$e_{ij} = r_i p_{ij}.$$

By computing edge fluxes, NetWalk can be utilised to postulate not only genes which are associated with a given phenotype, but also interactions, permitting this methodology a strong insight into the effects of network rewiring between two different conditions. Indeed the methodology was employed to investigate network rewiring in MCF7 breast cancer cells in response to various doses of the chemotherapeutic drug doxorubicin and revealed a switching behaviour in the p53 network controlling apoptosis and cell cycle arrest [16].

Other random walk based approaches include *Expression Quantitative trait loci Electrical Diagrams* (eQED), a methodology developed in [50] and applied in [15] to identify pathways dysregulated in brain cancer. The methodology integrates both expression data and SNP data with an experimentally verified network of biological interactions. The authors first identify a set of genes, which are differentially expressed in patients with the brain cancer Glioblastoma multiforme compared to healthy controls, defining this set of genes as the phenotypic layer of the network. Genes identified by SNP assays as holding genomic aberrations driving this cancer phenotype are also identified and are referred to as the causal layer of the network. The algorithm proceeds by utilising a random walk based analogy between signal transfer through a biological network and the flow of current in an electrical circuit. Edges in the network are assigned a conductance, based upon the expression correlation between the vertices they connect and the genes in the phenotypic layer of the network. The methodology then proceeds to investigate signal transfer in the network from the causal layer to the phenotypic layer by computing the currents along edges and the voltages at vertices, which can be found from considerations of Ohm's and Kirchoff's Laws. This methodology was thus capable of tracing paths of dysregulated

signalling which were driving the cancer phenotype and thus may represent relevant therapeutic targets, however, the daunting size of the causal layer made prioritisation of these pathways difficult.

In addition to methodologies like NetWalk and eQED, which focus on the local level of individual genes and pathways, network rewiring methodologies based on random walkers have also benefited from a more global, systems perspective. Methodologies of this flavour include the network entropy analysis introduced in [17] and explored in greater detail in [23] to identify systems properties of biological signalling in cancer. This methodology integrates gene expression data with an experimentally verified network of protein-protein interactions by weighing each edge (i, j) in the network with the correlation in gene expression between the genes i and j, across samples corresponding to a given phenotype. This edge weighting corresponds to a sparsified correlation matrix for each phenotype, which is then row normalised to create a stochastic matrix for each phenotype (one for healthy and one for cancerous tissue). This stochastic matrix: $P = (p_{ij})_{ij=1}^{V}$ (where V is the number of vertices in the network), describes the transition probabilities of a random walker in the network. Rather than investigate the stationary distribution of this walker, like NetWalk and NetRank, however, the next stage of this methodology considers the rows of the stochastic matrix P. The i^{th} row of P describes the probabilities that a walker starting at vertex i will transition to a given neighbouring vertex in the network. By considering the entropy of the i^{th} row of P, given by

$$S_i = -\frac{1}{\log deg(i)} \sum_{k \in \mathcal{N}_i} p_{ik} \log p_{ik},$$

one is able to determine the promiscuity of signalling from vertex i. If i has no preference in which of its neighbours it will send the walker to, then S_i will be 1, if i deterministically sends the walker to a given vertex then S_i will be zero.

This methodology was employed to compare metastatic and non-metastatic breast cancer [17]. By integrating the two cancer phenotypes separately with a protein interaction network and computing a vector of entropies for each, it was found that genes in metastatic breast cancer have significantly higher entropies than non-metastatic breast cancer. Moreover, by considering expression data from a large number of cancerous and healthy tissue samples, it was demonstrated that the average entropy of genes in cancerous tissue was higher than healthy tissue, implying that network entropy increases as cancer progresses [23]. Intriguingly, the entropy of oncogenes was lower in cancerous tissue versus healthy tissue and the reverse was true of tumour suppressors. This demonstrated that a rewiring from a high entropy state to a lower state, is indicative of activation of a gene in a pathway and revealed the power of the entropy based methodology in pinpointing critical

coordinators of a phenotype.
The experimental demonstration that biological networks were rewiring thus led to the birth of many new methodologies centred on comparing network structure and dynamics. Many powerful differential network algorithms are based on random walks on weighted networks, which represents a powerful framework for the integration of gene expression data with interaction data for the inference of phenotype level differences. Methodologies such as NetWalk and eQED provide a glimpse into differences between intracellular interactions between phenotypes, but the general principles and dynamic nature of network rewiring are still unclear. Network entropy, a measure of pathway promiscuity appears to show a clear directed increased as healthy tissue became cancerous and subsequently metastatic, indicating that disorder is a global property of protein interaction signalling that dynamically increases in cancer. This result is important as it suggests that systems properties may underlie network rewiring in pathology, permitting us to posit rules and laws regarding signalling changes that may provide the basis of a strong mechanistic understanding of complex pathology.

1.1.4 Perspectives: Relevance for our research

Our work focuses on the development of network theoretic tools for the analysis of complex disease with the ultimate goal being the identification and validation of novel drug targets.
It is clear that a critical first consideration is the selection of the most suitable interaction network. The work in this thesis will consider heterogeneous pathologies, described by both time course and steady state gene expression data sets of varying sample size and our aim is to develop methodologies which are relevant to all these scenarios. We have seen, however, that the main classes of network inference algorithm are highly data dependent and would likely provide very different results on different data sets. We thus will opt to use manually curated, experimentally verified interaction networks. We will use gene expression data to weigh our interactome and thus impose sample specificity. In addition, we will use *Gene Ontology* (GO) annotations to delete spurious interactions between proteins which are unlikely to co-localise, thus deleting false positives introduced by Y2H protocols.
The next important consideration for the the development of network theoretic tools is whether to consider single phenotype or differential networks. We have seen that both approaches have proved informative but that the differential network approach is likely to capture more biologically relevant information. Of such network rewiring approaches,

it is clear that random walk theory and stochastic analysis are powerful tools for modelling network traffic, with methodologies such as NetWalk and eQED elucidating novel differential interactions between phenotypes.

It is evident that many methodologies are emerging in this field, yet it is unclear which are superior for given applications. It seems that a reason for this may be that most current methodologies are based upon the identification of local network alterations, which are not scalable to a global property, *e.g.*, visitation frequency of a vertex in a weighted random walk [16]. To me this represents a problem; it is clear from experimental evidence that network rewiring is a global phenomenon, with thousands of genes and interactions modified by a single perturbation (*e.g.*, administration of MMS to yeast [20]). Moreover, biological networks are highly sophisticated dynamical systems, which have evolved to orchestrate tightly co-ordinated responses. It does not seem unlikely therefore, that there may be global properties underlying network rewiring, *i.e.*, only certain trajectories in network rewiring state space may be admissible. Moreover, one may find that global properties of network rewiring provide an ordering for the evolution of network traffic and topology during descent into pathology, the progression of a complex healthy biological process (such as cell differentiation), or following multi-drug administration.

Were such global properties to be uncovered a natural next stage would be the development of local analogues, to identify key features (*e.g.*, genes or pathways), which are driving the global shift. These may represent the most sensible targets for modifying a biological network, as their modification will have the greatest potential to affect global network rewiring.

We will thus focus on the development of random walk based network rewiring methodologies, which aim to characterise biologically relevant global properties of network rewiring that are scalable to local properties, for the subsequent identification of drug targets.

1.2 Mathematical approaches to understanding cell differentiation

In the previous section we explored the realisation that global genetic interactions are highly dynamic, shifting enigmatically as a cell responds to perturbation to maintain homeostasis. This has raised the question: is there a regularity to this restructuring? More specifically, are there certain important, natural dynamical biological process in which network rewiring, as assessed by some global systems property, proceeds in a reproducible and regular fashion? Were this the case, not only would the systems properties of the biological process in question be elucidated, but one would have a metric of network rewiring with a solid biological interpretation. Assessing such a global systems measure

of network rewiring in pathological cells would provide a proxy for the progression of an important biological process in the disease state, providing novel insights into what may be driving the pathology. The choice of biological process to investigate in this case must be made carefully to ensure the relevance to disease.
Arguably a suitable natural biological processes is cell differentiation. Representing the concerted orchestration of the expression of a vast number of genes, cell differentiation results in the temporally ordered convergence of form and function as a pluripotent cell moulds its way towards a terminal cell fate. Such dynamic modification suggests that this process is an excellent candidate for the identification of a systems property of network rewiring. Moreover, cell differentiation is frequently altered in complex diseases, most notably cancer [51] and developmental diseases such as FSHD [52], making it an informative process to measure from the perspective of subsequent analysis of pathomechanims.
Following a stem cell on its journey through the mysterious landscape of differentiation, and the precise measurement of cell potency, are problems that have plagued systems biology for more than 50 years [53]. Over time, however, general principles of what characterises a stem cell and how these alter over a differentiation trajectory have emerged. In what follows we review the history of this field. First we examine the experimental approaches to understanding the molecular circuitry governing cell fate committal, before exploring the more theoretical, systems approaches to assessing pluripotency, culminating with the most recent statistical mechanical concepts of MacArthur and Lemischka [54]. We close as above with some perspectives of relevance to our research goals. Most notably we express the notion that a network rewiring methodology may be developed on entropic principles to assess the validity of the statistical mechanical approach to cell pluripotency. This approach may form the basis of a methodology of the flavour we outlined above, namely one which aims at identifying global properties of network rewiring which may be scaled to more local ones to identify driver features.

1.2.1 Experimental approaches to quantifying cell potency: some limitations

During development pluripotent stem cells are stimulated to undergo complex and temporally ordered transcriptional programmes, resulting in unidirectional functional and morphological changes and culminating in the generation of a plethora of diverse terminally differentiated cells, each specialised for a given role within the organism. Understanding and control of this process is an exciting and active area of research, which may prove critical to the understanding and treatment of developmental pathology with disease modelling [55] and regenerative medicine [56]. Moreover, cell differentiation is

altered and possibly reversed in cancer and elucidation of the principles governing the directionality of the process may shed light on oncogenesis [57].

A critical question in stem cell biology, and one that has profound implications for the clinical applications of the discipline is: How do we measure cell potency accurately enough to track a cell through a differentiation programme?

A pluripotent stem cell is currently most rigorously defined by its ability to form a teratoma containing all three germ layers in immune-deficient mice [58]. However, this qualitative, binary definition does not provide us with a metric with which to track cell potency. Without such a quantitative measure, the translation of stem cell medicine to the clinic must overcome some daunting hurdles [59].

Stem cell therapies typically initially require the isolation and propagation of stem cells. Following isolation, these cells may either be coerced to differentiate into a particular cell lineage required for therapy, via the use of growth factors, before transplantation [60], or simply introduced into the tissue to be repaired, where microenvironmental cues determine cell fate [61]. Pluripotent stem cells, such as *embryonic stem* (ES) cells, however, in addition to having the ability to generate all three germ layers also have a high proliferative capacity and are potentially tumorigenic [62]. We currently lack understanding of how a stem cell may differentiate into a tumour, or be safely guided towards a given tissue to be repaired. This represents a large barrier to the clinical applications of regenerative medicine [51]. The construction of an intuitive measure which can reliably distinguish pluripotent cells, cancerous tissue and healthy terminally differentiated cells would thus represent a critical step towards the development of strategies for regenerative medicine, which may avoid oncogenic complications. In addition, such a measure, if based on theoretical principles, would provide a conceptual step in the understanding of de-differentiation leading to oncogenesis, and would be able to empirically assess recent hypotheses regarding *cancer stem cells* (CSCs) and tumour cell plasticity [63, 64].

Stem cell therapy has long been plagued by the issue of immune rejection of transplanted ES cells. Whilst strategies to induce tolerance have been suggested [65], the most exciting potential solution is the use of *induced pluripotent stem cells* (iPSCs) which were first generated by Takahashi and Yamanaka in 2006 [66]. The two authors demonstrated that by the induction of just four factors, *Oct3/4*, *Sox2*, *c-Myc*, and *Klf4*, mouse fibroblasts could be converted into pluripotent cells capable of generating all three germ layers. This study had great implications for the overcoming of immune rejection of transplanted cells for regenerative medicine, as iPSCs can be derived from the skin cells of the patient to be treated, and thus are a perfect genetic match. However, whilst certain studies have demonstrated successful acceptance of transplanted iPSCs and derived cells in mice

[67], this finding is not universal and controversies over immune rejection of these cells is widespread [68]. Moreover, clear molecular differences between iPSCs and ES cells have been documented, and functional differences regarding their in vitro differentiation capacity currently limit clinical applications of iPSCs [69]. The molecular underpinnings of these functional differences are poorly understood, however, and a true understanding would require the ability to track cell potency as iPSCs and ES cells differentiate into a specified lineage, for direct comparison. Despite these difficulties iPSCs have found a powerful application in the modelling of complex diseases, providing a potentially limitless source of (often difficult to biopsy) cells for in vitro studies and the screening of therapeutics [55, 70, 71]. Such disease models often require the carefully guided differentiation of pluripotent stem cells into the pathological tissue of interest, however, without a metric for cell potency, deciding that a cell has reached a particular fate relies on qualitative assessment of phenotype [55]. Such a subjective and non-quantitative measure for obtaining disease model cells may result in large differences on the molecular level, between the iPSC derived cell and the true cell. This has the potential to render conclusions derived from pluripotent stem cell models inapplicable in a clinical setting, and emphasises the need for a more quantitative measure of cell potency, derived from a global understanding of the mechanisms of cell differentiation.

Constructing a quantitative definition of pluripotency which can follow a cell through any lineage, however, appears to be a highly non-trivial pursuit. It was initially proposed that stem cells expressed a unique set of molecular markers which would allow their robust identification from committed cells. Consequentially, several microarray studies attempted to extract a molecular signature of 'stemness' [72, 73]. However, the signatures identified in these early studies were refuted, and a reproducible microarray assessed gene expression based classifier for identifying pluripotent stem cells remained elusive [74, 75]. Genetic studies proved more useful in the identification of genes essential for the establishment and maintenance of pluripotency, such as *Oct4* and *Nanog* [76]; however the expression of even these key genes appeared to be highly variable among pluripotent cell populations [54]. Functional studies were able to demonstrate the ability of certain pluripotency factors to determine the early stages of differentiation, such as bimodality in *Nanog* expression dictating either the asymmetric or symmetric division of pluripotent cells [77]. In addition considerable efforts have been made in identifying the regulatory networks essential for the maintenance of pluripotency. These include the work of Wang *et al.* [78], in which mass spectrometry was used to quantitatively asses the protein-protein interactions with *Nanog* in ES cells to identify regulatory modules required for the maintenance of pluripotency. In addition Muller *et al.* published two extensive microarray studies [59, 79]

profiling gene expression in a considerable number of pluripotent and multipotent stem cell lines. In their first paper Muller and co-authors utilised graph theoretic techniques to derive a putative network of interactions required for pluripotency, which they named *PluriNet* [79]. In their second study the authors utilised a machine learning algorithm to classify microarray samples as pluripotent or not [59]. Their resulting bioinformatic assay, dubbed *PluriTest*, was intended to be a quantitative, standardised alternative to the teratoma assay and displayed a high sensitivity and specificity. However, beyond these early stages of loss of pluripotency, commitment to different cell lineages requires such a highly diverse set of pathway activation and gene expression programmes, depending on terminal fate, that consideration of the expression of specified gene sets is unlikely to provide a measure of potency that is robust across all lineages. Rather, to derive such a measure one must consider not which genes are expressed, but how the genes are expressed in concert; one must therefore consider the systems biology of cell differentiation.

1.2.2 Systems biology of cell differentiation

1.2.2.1 Waddington's Landscape The study of the systems biology of cell differentiation has a rich history, of the early approaches the most successful was doubtless the epigenetic landscape concept introduced by C. H. Waddington [53] (Fig 1.3). Within this framework, the cell is represented by a ball on a sloping landscape, pluripotent stem cells occupy high positions, capable of rolling down branching paths to eventually rest in one of multiple valleys, each representing a differentiated cell fate. The structure of the landscape was postulated to be governed by interactions between genes and proteins, though the precise relationship was not elaborated upon. However, by dictating a potential gradient from stem cell to committed cell, and hills separating terminal valleys, Waddington incorporated a low probability of de/trans-differentiation directly into his model. However, the structure of the landscape did not disallow de/trans-differentiation, it simply suggested that such phenomena would require large perturbations to overcome the potential wells of committed cells. Indeed this concept was validated by the generation of iPSCs, in which the perturbation of the cell by the introduction of pluripotency factors was sufficient to induce de-differentiation.

Though it was intended as a qualitative description, Waddington's landscape became an intuitive framework for a more quantitative dynamical systems analysis of cell differentiation, where it was seen as the gene expression phase space [80]. Under a dynamical systems model, interactions between genes can be represented as systems of differential equations, the nullclines of which can be used to construct a phase space describing the

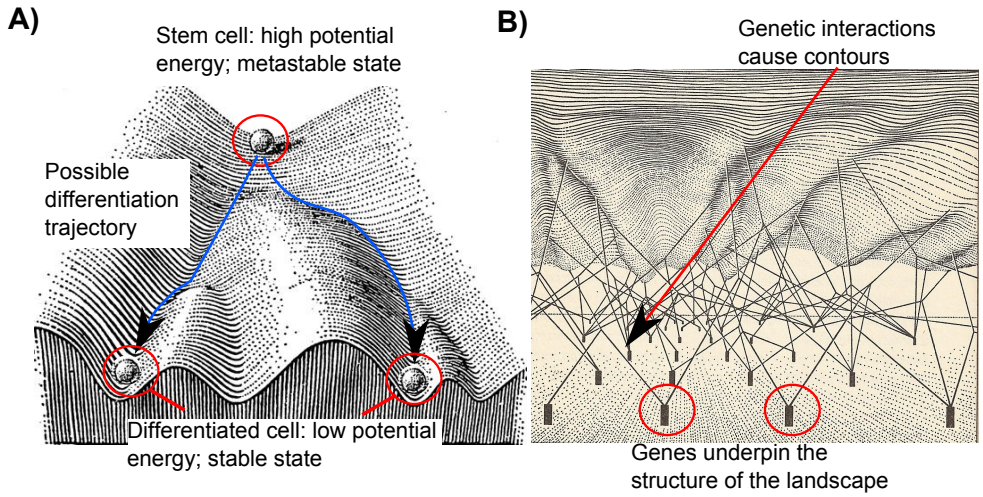

Figure 1.3: **Waddington's Landscape.** (A) Waddington's original sketch of the differentiation landscape, stem cells occupy positions of high potential energy, and differentiate by losing this potential. (B) A later sketch by Waddington hypothesises that the structure of the landscape is underpinned by genetic interactions. Adapted from [53].

evolution of a gene expression trajectory emanating from any starting state [81]. The valleys representing terminally differentiated cells find a natural interpretation as attractors of the system, whilst meta-stable states, from which a trajectory may escape given a sufficient perturbation and converge to a stable attractor, can be interpreted as stem cell states [82]. This notion of high dimensional gene expression phase space attractors representing committed cell fates, dubbed the attractor hypothesis, has been supported by direct experimental evidence. Huang *et al.* [83] generated microarray time courses of two independent stimulations of human HL60 promyelocytic progenitor cells differentiating into neutrophils. One time course represented differentiation stimulated by *all-trans retinoic acid* (ATRA) and the other by *dimethylsulphoxide* (DMSO). Utilising principal component analysis, the authors demonstrated that whilst the gene expression trajectories of the two stimulations initially diverged (representing different perturbations from the progenitor cell state) they subsequently converged to the same stable terminal state rather than exploring the state space. This is the same behaviour as trajectories in a dynamical system converging to an attractor in phase space. Whilst this result is convincing, alternatives to the attractor hypothesis have been suggested, such as the work of Mar and Quackenbush [84], which considered cell differentiation as a superposition of transient and core processes. In addition, a new model has recently been proposed for cell

fate committal in an attempt to adjust Waddington's theory to account for feasibility of de/trans-differentiation now possible in vitro. Dubbed the epigenetic disc [85], this new model discards the potential slope of Waddington in favour of flat landscape, which may be tilted by the introduction of various factors to induce trajectories from one cell fate to another. In the context of this new model, the clinically important question of how to measure cell potency, *i.e.*, measuring the height in Waddington's landscape, is replaced by measuring the gradient of the disc during transition between cell states.

1.2.2.2 Small regulatory networks and binary decisions Given the huge complexity of genetic interactions, many dynamical systems studies aiming for high resolution of the molecular mechanisms of cell differentiation have focused on small regulatory circuits. Of these perhaps the most well studied is a mutually repressive pair of auto-activating transcription factors [86, 87]. This simple genetic circuit, common in the specification of cell fate, admits a dynamical systems model with a phase space which is highly intuitive under the attractor hypothesis (Fig 1.4). The system has two stable attractors, each corresponding to high expression of one transcription factor and no expression of the other, in addition, there is a meta-stable state corresponding to the moderate expression of both transcription factors. Clearly a small perturbation from the meta-stable state, in which the quantity of one transcription factor is increased above the other, initiates a convergent trajectory towards the attractor state in which only the perturbed factor is present. Such a regulatory circuit provides an elegant means for binary decision making, as a cell which over-expresses one factor commits to a given cell fate whilst over-expression of the other factor initiates a different, mutually exclusive cell fate. The most well studied example of this regulatory circuit in the determination of cell fate, is the case of *GATA1* and *PU.1* in the specification of erythroid and myelomonocytic lineages [88]. In progenitor cells both *GATA1* and *PU.1* are expressed at low levels, however, following introduction of a stimuli which perturbs the levels of these factors, the cell will either converge to an erythroid state, in which $GATA1 >> PU.1$ or a myelomonocytic state where $PU.1 >> GATA1$. This notion of expression reversal in dictating cell fate was explored in greater detail in a recent study by Heinaniemi *et al.* [89], where the authors investigated gene expression data from 166 cell types to identify gene pairs which displayed opposite expression levels between lineages. This exploratory data analysis identified a number of candidate circuits involved in the control of cell fate and pluripotency that exhibited a similar structure to the classic *GATA1-PU.1* circuit.

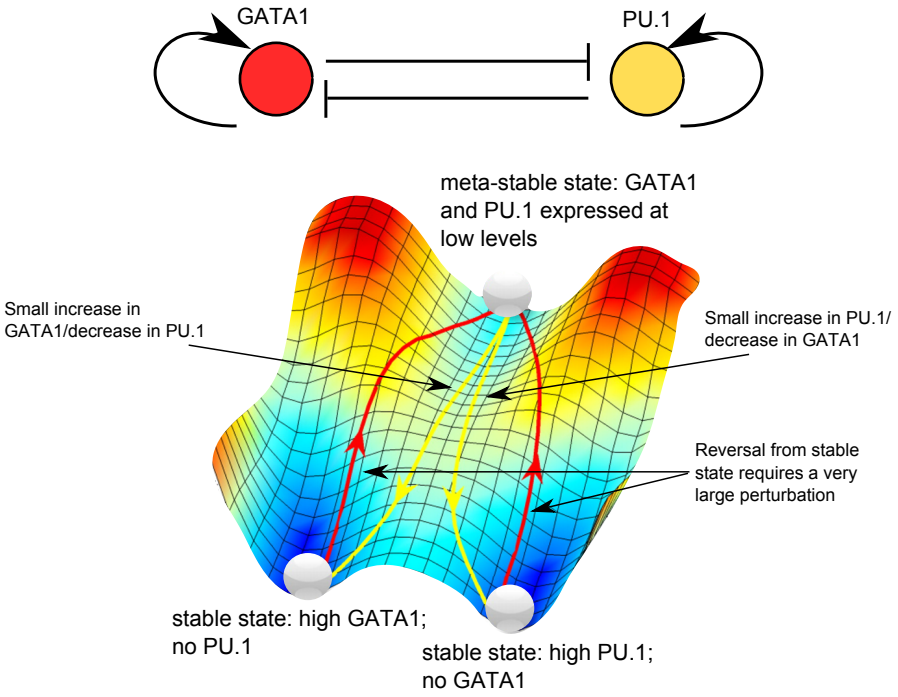

Figure 1.4: **Small regulatory circuits for binary decisions.** A pair of auto-activating mutually repressive transcription factors such as *GATA1* and *PU.1* are shown, alongside a phase space evaluated by Wang *et al.* [90]. We see that the circuit results in a meta-stable state, which can be easily perturbed to one of two stable states corresponding to high expression of only one of the transcription factors. This small circuit can thus be used to make binary decisions. Adapted from [90].

1.2.2.3 Stochasticity and cell fate It is well known that gene expression is an inherently noisy process [91], with the expression levels of genes and proteins subject to high degrees of variability, in a manner which depends upon a number of factors, including, crucially, the network architecture of genetic interactions, which can act to supress or enhance stochastic fluctuations [92, 93]. Given that cell differentiation can proceed along independent lineages as the result of small perturbations in the levels of certain transcription factors, it was suggested that stochasticity may play an important role in the determination of cell fate [21, 94]. Indeed it has been proposed that noise is essential in the selection of cell lineage [95], for example in *Drosophila melanogaster* where neural precursors arise consequential of stochastic fluctuations in protein levels in a homogeneous population of cells [96]. Rather intriguingly, it has also been demonstrated that stem cells display a considerable, dynamic variability in their expression levels. This variability has been interpreted as noise driven single cell transitions between multiple meta-stable attractor states corresponding to the same stem cell, but with different biases for lineage committal [97]. This variability of single stem cells, is counteracted by a robustness at the population level for the proportions of cells assigned to each meta-stable sub-state, which appears tightly regulated [21]. The ability of stem cells to exhibit such noisy dynamic variability in their expression profiles, allows them to maintain the option of committing to a multitude of cell fates. This is in stark contrast to committed cells which must maintain robust activation of many pathways to prevent de/trans-differentiation and ensure functionality. Thus we have stumbled upon a hint as to a systems property underlying cell potency which may be utilised to construct a global measure of differentiation potential: gene expression in stem cells is highly variable and noisy and this variability decreases as cells differentiate [58].

1.2.2.4 Quantifying randomness in stem cell gene expression Recently, several attempts have been made to quantify the observation that stem cells exhibit a more variable and dynamic expression profile and construct a measure of pluripotency based upon global analysis of stem cell and differentiating cell molecular profiles. To date, however, these works have remained largely hypothesis driven and theoretical, with little attempt to apply the measures to experimental data. These hypotheses include the dynamical systems based approach of Furusawa and Kaneko [98], in which the authors postulated, based on the observation of stem cell heterogeneity, that pluripotent stem cells would exhibit a highly chaotic gene expression dynamic. This chaotic attractor of stem cells was postulated to be meta-stable, with stochastic fluctuations in gene expression permitting

divergence to new attractor states representing more differentiated cells, which would be less irregular, though still chaotic. This process was postulated to continue until a stable attractor was eventually reached, representing the terminally differentiated cell. The authors provided a proof of principle of their hypothesis via the use of genetic algorithms to construct gene regulatory networks evolved under a fitness function aimed to maximise cell type heterogeneity (number of distinct terminal states to the expression dynamic). These simulations demonstrated that the proposed sequence of chaotic attractors was the most likely outcome of the genetic algorithm. While this result is not without interest, in that it aims to quantify cell potency through the 'chaocity' of expression dynamics, the immediacy of application to measuring cell potency experimentally is not clear, though an approach based on *fluorescence-activated cell sorting* (FACS) was proposed.
A second hypothesis was suggested very recently by MacArthur and Lemischka [54] and was based upon statistical mechanics, postulating the concept that entropic measures may be suitable to quantify the uncertainty in gene expression in pluripotent stem cells. The authors suggested that if one could construct a probability mass function, describing gene expression in a given cell population, then one could measure pluripotency as the entropy of this distribution. The logic behind this theory being that stem cells have highly variable gene expression profiles and thus will display a more uniform probability distribution of gene expression than committed cells. Whilst this theory is elegant and would indeed provide a quantitative measure of cell potency, the authors deliberately avoid any suggestion as to how to estimate the probability mass function essential for computation of the entropy, making application elusive.

1.2.3 Perspectives: Relevance for our research

The aim of our research is to develop and apply network theoretic tools for the analysis of complex diseases, with the objective of postulating and validating novel drug targets. In the first review in this section, on network biology, we identified a gap in the literature for a methodology which utilises scalable global measures to characterise systems properties of biological network rewiring and identify key drivers of these global alterations.
Developing such a methodology requires careful consideration. Primarily one needs a hypothesis pertaining to a global systems property of network rewiring during an easily measurable dynamical biological process. Moreover, this hypothesis needs to be expressible in mathematical terms, as a global network measure, which can be scaled to a local measure. Given these criteria one must then use experimental data to test the validity of the hypothesis. Only if the hypothesis is confirmed can we proceed to develop our

network rewiring methodology in a manner that is biologically realistic.
Given the pathologies we are investigating in this thesis (namely cancer and FSHD) display defects in cell differentiation, we believe this natural biological process to be highly suitable, as a measurable phenomenon which may be quantified by a property of global network rewiring. Recent findings in stem cell biology have led to the postulate that gene expression uncertainty is high in stem cells and decreases throughout differentiation. We have also seen that pluripotent cells have the possibility to activate of a large number of pathways to commit to various lineages, whilst committed cells must maintain fixed activation of pathways specific to their fate (*e.g.*, the *GATA1-PU.1* regulatory circuit in hematopoietic cells). These observations leads us to postulate that intra-cellular networks are rewiring during differentiation from a highly promiscuous, non-preferential, meta-stable state in stem cells to a more deterministic state in committed cells. The promiscuity of network signalling can be measured by adapting the network entropy measure constructed in [17] and explored in the first introductory section.
Thus in this section we have explored cell differentiation as a complex biological process of great relevance to our complex diseases of interest. We have seen how pluripotency may be considered as a state of highly disordered promiscuous signalling, and proposed entropy as a measure which may associate with differentiation potential. Elegantly this measure can be applied both globally and locally, precisely as required for a methodology of the flavour described in the first introduction. We will thus explore in this thesis a form of network entropy in the context of a measure of cell potency.

1.3 The pathophysiology of breast cancer

In this thesis we aim to investigate complex pathologies, in which cell differentiation and development play a critical part. Of such pathologies, cancer is arguably the most well know and exalts the most severe emotional response. Though we will touch on a range of epithelial cancers, our main focus will be breast cancer. A complete introduction to all tissue specific malignancies considered is thus beyond the scope of this work. Hence in this introductory section we will focus solely on neoplasms of the breast, as a representative member of the class of oncogenic pathologies we will work on.
Breast cancer is a highly complex pathology, requiring the involvement of a diverse set on intra and extracellular interactions, which conspire to drive the progression of neoplastic malignancy. The condition is characterised by a considerable heterogeneity in clinical presentation, morphology (as assessed by histology) and in molecular mechanisms, which differ vastly between patients. In addition to being a highly sophisticated condition, breast cancer is also the most common cancer among women (30.7% of female cancer cases in

the UK in 2011; ONS Cancer Registration Statistics [99]), and deaths from breast cancer represent the second most common cause of death among women in the UK (15% of all female deaths from cancer). These statistics highlight the importance of the development of treatments and therapies for this highly complex condition.
A key aim of our thesis is the application of network theoretic tools to understand breast cancer molecular mechanisms. To this end it is important to consider a number of factors. A first consideration is how the condition is diagnosed and the nature of the sample one can obtain from patients for investigation. A second consideration is the sub-classification of the pathology into phenotypically distinct subtypes, and whether the understanding of differences between subtypes can be improved by the application of network theoretic tools. Given our interest in cell differentiation and our postulate that entropic measures may prove powerful metrics of cell potency, we should also consider the notion of stemness in malignancy, generating hypotheses around how an entropic measure of differentiation potential may prove useful in understanding this concept.
In what follows, therefore, we review the pathobiology of human breast cancer, beginning briefly with diagnosis via imaging and histology, and moving on to subtype classification, initially by a handful of molecular markers and subsequently by genome wide expression profiling. We then explore the cancer stem cell (CSC) hypothesis, noting potential applications of a cell differentiation based network rewiring approach. We close, as usual, with some perspectives of the relevance of this review to the work pursued in our thesis. In particular we note how an entropic approach, explored at the end of the previous introductory section on stem cell biology, may prove useful in the identification of drug targets for certain breast cancer phenotypes.

1.3.1 Breast cancer diagnosis: imaging to biopsy

In the data driven analysis of any pathology, it is important to first consider where one's data comes from. Typically, data is acquired following the completion of diagnostic testing and the limitations and epidemiological non-uniformity of such testing may introduce bias to one's sample. To this end we here consider breast cancer diagnosis via imaging. Many countries run national breast cancer screening programmes for woman at increased risk of developing the condition; one third of breast cancer patients are diagnosed following such screens [100]. There has been much debate over the costs and benefits of screening programs. The major benefit of screening is believed to be in the early detection of breast cancer, which permits therapeutic intervention at a stage where the cancer is most treatable. Costs of screening are generally attributed to over-diagnosis, in which cancers are

detected which would otherwise not be noticed during the lifetime of the patient, resulting in unnecessary surgery as well as detrimental psychological effects. A recent independent study analysing the costs and benefits of breast cancer screening in the UK, based upon a meta-analysis of 11 randomised control trials following patients aged between 50 and 70 for at least 13 years, concluded that patients invited to undergo screening benefited from a 20% reduction in breast cancer mortality [101]. Though the authors of the study commented that the assessment of over diagnosis was somewhat subjective, and that no 20 year follow up study into its adverse effects existed, they concluded that the benefits of screening outweighed the costs.

It must be noted, however, that the benefits of screening are highly age dependant. Indeed the reduction in mortality rate due to screening among women aged 40-49 years (15%) is nearly 5 times less than for women aged 60-69 years [102].

Mammography is considered the gold standard method for breast cancer screening and diagnostic imaging, and employs low energy X-rays to image the breast for subsequent determination of tumour masses [103]. Though this low dosage of radiation is unlikely to be carcinogenic in post-menopausal women, it was demonstrated that in women younger than 40 the beneficial effect of biennial mammography from early detection and treatment of breast cancer, was offset by the oncogenic effect of exposure to radiation [104]. In addition to the increased risk from radiation exposure, conventional screen film mammography images are often difficult to interpret and it is estimated that radiologists fail to detect between 10% and 30% of tumours, with this error rate rising in dense breast tissue [105], where the risk of developing malignancy is four to six times higher [106].

Despite these limitations, imaging is a critical part of the breast cancer treatment pathway. Providing the essential initial insights required for preliminary diagnosis, as well as a powerful guide for surgical intervention to remove tumour masses.

Once a tumour is diagnosed via imaging, it is biopsied to confirm. This biopsy can be investigated via a number of means, including histologically and molecularly, to provide clinical information which can subtype the tumour and guide a suitable treatment course, often beginning with surgical excision. We will, in this thesis, consider gene or protein expression data from bulk tumours. The data we will consider typically corresponds to the biopsy or surgery stage in progression. Whilst this is likely to provide a representative subset of most patients, it must be made clear that the nature of image based diagnosis will somewhat skew our sample. We will likely not detect dense breast tumours at as early a stage as tumours that arise in less dense breast. Similarly, we will likely detect tumours earlier in post-menopausal women, and women with certain genetic dispositions to breast cancer who undergo regular mammography screening. Women who are not screened as

regularly will likely present at a later stage. It is important to be aware of these biases before considering the data.

1.3.2 Histology: subtyping on cellular composition

As mentioned, after detection by imaging of the breast, an abnormality is generally biopsied whereupon the tumour sample is examined and classified. This classification allows us to consider breast cancer as an array of pathologies, each with their own rates of progression and responses to therapy. Understanding this subtyping will allow us to tailor our methodologies and investigations to these distinct sub-diseases.

The vast heterogeneity of breast cancer means that classification is a complex procedure. The first stage is generally histological classification, the consideration of tumour cell morphology to divide tumours into distinct groups. Briefly, biopsied tumours are sliced in a cruciate manner, in the fresh state (not frozen) immediately after resection and fixed in 10% phosphate buffered formalin. Sections sufficiently thin for the elucidation of nuclear detail (4-6μm) are then cut and stained with *haematoxylin and eosin* (H&E) and analysed by microscopy [107].

The two most important histological classifications from a prognostic perspective are histological grade and histological type [108]. Histological grade is a measure of cell anaplasia (level of de-differentiation) and is assessed by the combined scoring of three features: tubule formation, nuclear pleomorphism and mitotic counts [107]. Assessment of these three features is generally performed via a slightly modified Nottingham grading method [109]. Briefly, nuclear pleomorphism is assessed on a scale from 1 to 3, with low scoring samples displaying small, low variability sized nuclei, with inconspicuous or no visible nucleoli; high scoring tumours in contrast have large variable sized nuclei, which are often vesicular and irregularly shaped. Tubule formation is again scored from 1 to 3, with low scoring tumours having $> 75\%$ of the sample comprising tubular structures; high scoring tumours are dominated by vacuolated cells and solid sheets of cells. Mitotic counts are again scored from 1 to 3 with high scoring samples having a large number of mitotic cells per sample area. These three scores are then combined to provide a grade from 1 to 3 for each sample. It has been demonstrated that low grade tumours have a significantly more favourable outcome than high grade tumours and thus the assessment of histological grade is an important stage in clinical diagnosis of breast cancer [107].

Histological type is a classification of tumour samples based upon the similarity of tumour cells to common differentiated cell types. All breast cancer tumours which arise from the mammary epithelium are by definition, based on tissue of origin, adenocarcinomas [110]. Among observed neoplasms of the breast histologists have identified one

common type and 17 special types [111], which are characterised by certain morphologies and which comprise about 25% of all breast tumours [108]. The 75% of common breast tumours are classified as *invasive ductal carcinomas of no special type* (IDC-NST) [108]. Given that their definition is simply that over 50% of their observed morphology is not of a special type [108], IDC-NST tumours exhibit a great variability in their morphology, displaying diversity in nuclear pleomorphy, cell dispersal and glandular formation [112]. Consequentially the *World Health Organisation* (WHO) sub-classified IDC-NST tumours into 4 distinct subtypes based on observed features: pleomorphic carcniomas, carcinomas with osteoclast-like giant cells, carcinomas with choriocarcinomatous features and carcinomas with melanotic features [111]. The proportions of the different histological subtypes found in a large breast cancer data sets are displayed in Fig 1.5A.

A tumour is said to be of pure special type if >90% of its morphology matches that of the considered type, if the tumour is a 50-90% match, it is referred to as a mixed type. Mixed types have been reported at a higher prevalence than pure special types (25.3% *vs.* 18.3%) [113].

Of the special histological types of breast cancer the most common are invasive lobular carcinomas which comprise 5-15% of all breast cancers [111]. Invasive lobular carcinomas are characterised by their migration pattern as single cells or in single file lines, often making diagnosis difficult as they do not always present as a lump [108]. Morphologically the cells of invasive lobular carcinomas are generally regular, small, round and with little cytoplasm [112]. Heterogeneity among invasive lobular carcinomas has led to their sub-classification into 4 distinct groups: classic, alveolar, solid and tubulolobular [113]. Alveolar lobular carcinoma presents with small aggregates of 20 or more cells [112]. Solid lobular carcinomas invade the surrounding stroma in sheets rather than single file lines [112], whilst tubulolobular carcinoma invades the stroma in cords which occasionally form small tubules [112]. Lobular carcinomas as a global group have a similar prognosis to IDC-NST tumours, however tubulolobular carcinoma holds a survival advantage over IDC-NST, whilst solid lobular carcinoma has the worst prognosis of the group [112]. Lobular tumours are also characterised by an increased tendency to metastasise to the gastrointestinal tract and gynaecological organs [108] and are less responsive to adjuvant chemotherapy than other breast cancer types [114]. It should be emphasised that although it was initially hypothesised that ductal and lobular carcinomas derived from ductal and lobular epithelium respectively, this hypothesis is poorly supported [112, 108].

Among the other special types, tubular and medullary tumours are the most common [108]. Tubular tumours are characterised by presence of infiltrative tubules extending into connective and adipose tissue and hold a survival advantage over IDC-NST [112].

Medullary tumours are characterised by interconnecting sheets of pleomorphic cells, lymphocyte infiltration and sharp borders. Medullary tumours also convey a survival advantage over IDC-NST of matching grade (generally grade 3) [112]. It should be noted that the fact that tumours of different histological type but of the same histological grade have different prognoses testifies to the orthogonality of prognostic implications for these two classifications. The remaining special types are considerably rarer and thus less is known concretely about associated survival, however as a group these rare special types also display superior prognosis to IDC-NST [112].
Thus histological classification has provided a tumour subtyping of prognostic relevance which in some cases can predict the efficacy of given treatments. These classifications however, whilst useful in this sense, do not provide any information as to the reason why certain tumours are more aggressive than others. Obtaining this information is key to the development of therapies tailored to combat the most aggressive forms of breast cancer and requires the consideration of molecular mechanisms.

1.3.3 Molecular subtyping: Hormone receptors and *HER2*

Whilst histological classification is routine and provides essential prognostic information about breast neoplasms it is performed by eye and hence is subjective. Moreover, it provides limited information as to molecular mechanisms underpinning the aggressive nature of certain subtypes. An understanding of such mechanisms can inform the development of targeted therapies. For our purposes, it is thus important to appraise the field of molecular subtyping of breast cancer.
Understanding possible molecular mechanisms of breast cancer has focused for many years on *hormone receptors* (HRs), of which the most important appear to the *oestrogen receptor* (ER) and the *progesterone receptor* (PR). Elucidating the involvement of HRs in breast cancer has been a long and difficult process and one which has not yet approached its conclusion. The first indication of their involvement came from a paper published in 1869 by G.T. Beatson [115]. In this article Beatson explained his ovarian theory of breast cancer which was based on observations of fat accumulation in the lactating breasts of rabbits following oophorectomy. Beatson also reported a case in which surgical removal of the ovarian tubes and ovaries coupled with thyroid tablets led to remission of breast cancer in a case that was considered inoperable. This finding was the first hint that endocrine therapy may be a suitable course of action in the treatment of certain breast cancers. However, it was clear that not all breast carcinomas responded well to such therapy. It was not until the characterisation of the ER and the development of sensitive

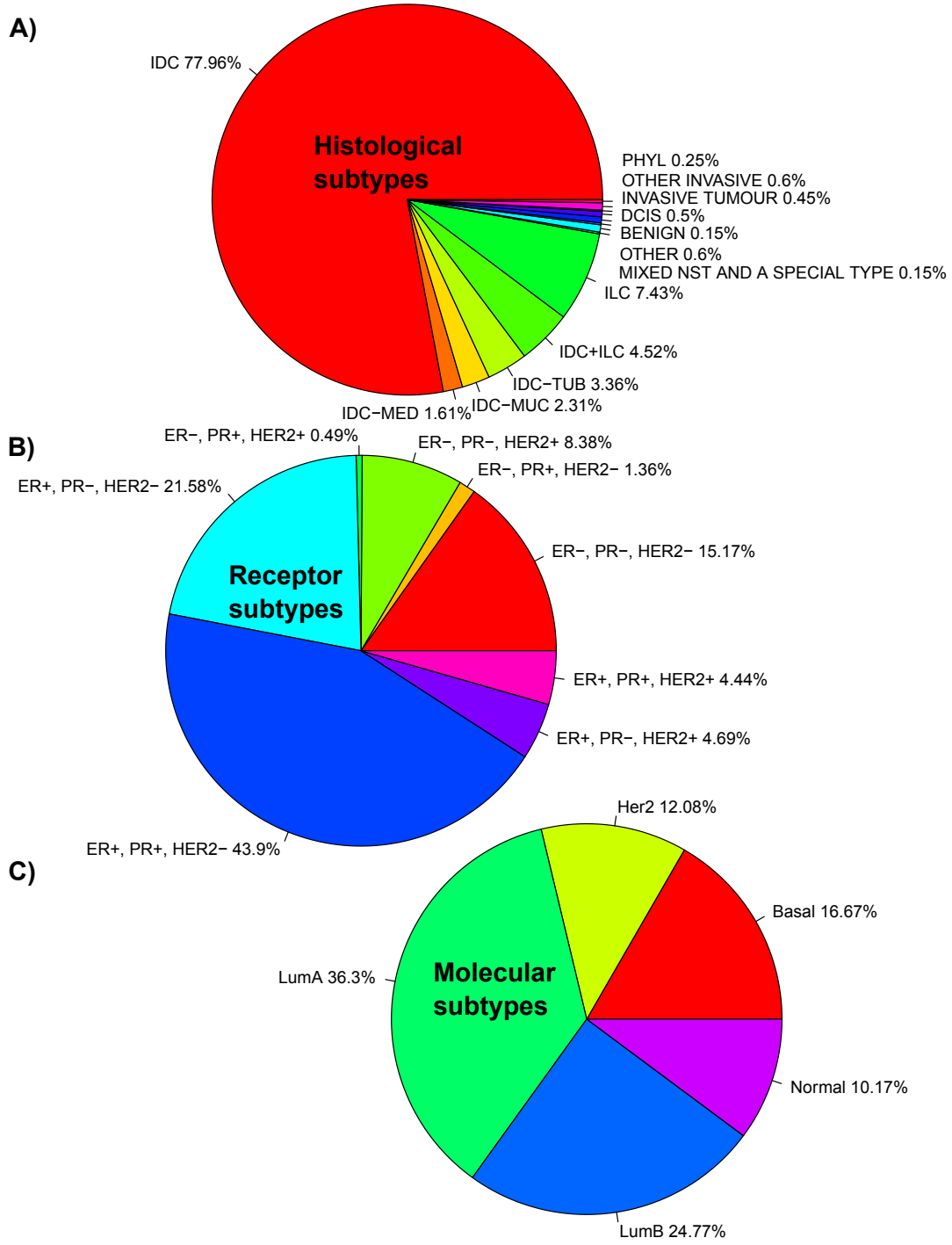

Figure 1.5: **Breast cancer subtype proportions.** Analysis of the METABRIC data set of 1980 breast cancers shows the proportions of (A) histological subtypes, (B) HR and *HER2* expression and (C) molecular subtypes.

assays to measure its concentration in tumours in the 1970s (most notably the dextran-coated charcoal assay [116]) that a rational understanding of which patients to treat using endocrine therapy was gained. The 1980s saw the development of monoclonal antibodies for the ER, which permitted the assessment of ER status by immunological techniques [117]. Refinement of these protocols and antibodies led to immunohistochemical assays, permitting the coupled assessment of ER status and cell morphology, which remains the gold standard assessment today.

Approximately two thirds of all breast cancers are assessed as ER positive [118]. ER is a member of the nuclear receptor family of ligand activated transcriptional modulators. Upon activation of ER by the binding of oestrogens, the protein is translocated to the nucleus where it is carefully directed by epigenetic mechanisms and cofactors to bind to target gene promoters and regulate transcript expression [119, 120]. ER orchestrates the expression of numerous genes involved in cell proliferation, angiogenesis, metastasis and inhibition of apoptosis [103]. Consequentially, inhibition of ER activity via various means has proved a successful endocrine therapy for many breast cancer patients. Among such therapies the most well-known and successful is tamoxifen [121], which antagonises the binding of oestrogens to ER and thus inhibit its activity. Aromatase inhibitors, which block oestrogen synthesis have also proved successful [119]. These therapies are essential to the treatment of breast cancer and it is estimated that approximately 400,000 women are alive today as a direct consequence of tamoxifen therapy [122]. However, it should be noted that the conventional 10 years of tamoxifen treatment is associated with an increased risk of endometrial cancer and pulmonary embolism [118].

Despite these advances, one third of breast cancers are ER negative and thus show no response to anti-oestrogen therapies. Moreover one third of ER positive cancers treated with tamoxifen will relapse within 15 years [123], implying that such cancers have an innate or acquired resistance to endocrine therapy. Consequently, further biomarkers and sub-classifications were required to improve targeted breast cancer therapies.

Many biomarkers were subsequently proposed, among them the most informative for clinical treatment were the PR and the *HER2* membrane receptor.

PR, like ER is a nuclear receptor and is found in the majority of breast cancers [124] but rarely seen outside of ER positive cases [103]. Progesterone (the PR agonist) is also known as the pregnancy hormone due to its increased expression and critical role in pregnancy. Moreover, it was demonstrated that the risk of developing PR positive breast cancer is substantially lower in individuals who have carried full-term pregnancies early in life [124]. Much like with ER positive cancers, antagonists were developed for the PR as a potential therapy for PR positive cancers. However, unlike tamoxifen these interven-

tions resulted in severe side effects, such as liver toxicity, and thus failed to succeed as therapies [124]. The prognostic implications of PR expression are debated [103], however it has been suggested that PR expression in ER positive breast cancer results in a better prognosis and a stronger response to tamoxifen therapy [125, 126].

The *HER2* membrane receptor is a member of the receptor tyrosine kinase family of *epidermal growth factor* (EGF) receptors. These receptors initiate the activation of signalling pathways that culminate in the up-regulation of gene sets involved in cell proliferation and which are anti-apoptotic [127]. The *HER2* gene was first identified as an oncogene through studies into chemically induced neuroblastoma in rat cell lines. It was demonstrated that genomic DNA isolated from such cells could induce oncogenic transformation when introduced into healthy mouse fibroblasts, and that this isolated DNA encoded the *HER2* gene [128]. Support for the oncogenic nature of *HER2* also came from studies into tumour viruses, where it was discovered that the avian erythroblastosis virus encoded a copy of a *HER2* related gene, which it retro-transposed into host cells to induce oncogenesis [110]. Subsequently it was found that many tumour-virus derived oncogenes were present in increased copy numbers in certain non-viral cancers, and notably the *HER2* amplicon was found amplified (almost exclusively [127]) in a subset of breast cancer samples [110]. It was then demonstrated that breast tumours displaying *HER2* amplification were of significantly worse prognosis than *HER2* negative tumours, and that the absence of this biomarker may in fact be more a powerful indicator of survival than HR status [129]. The *HER2* membrane receptor holds a constitutively active conformation, and its abundance due to over expression in *HER2* positive breast cancer is believed to lead to oncogenic pathway over-activation through dimerization with other EGR membrane receptors, most notably *HER3* [130].

About half of *HER2* positive tumours express either ER or PR, however, the levels of expression of the hormone receptors are considerably lower in *HER2* positive breast cancer than in *HER2* negative, resulting in a decreased efficacy of tamoxifen therapy in *HER2* positive ER positive breast cancers [127]. Thus in addition to being more aggressive, *HER2* positive breast cancers are also resistant to anti-oestrogen therapy. Consequentially new therapeutic strategies were required to tackle this breast cancer subtype.

Trastuzumab (also know as herceptin) is a monoclonal antibody directed against the extracellular domain of the *HER2* membrane receptor and is currently used as an effective treatment for *HER2* positive breast cancer [131]. Trastuzumab has also been reported to potentiate the effects of chemotherapy, making synergistic treatment a possible strong option [127]. The precise mechanism by which Trastuzumab combats *HER2* positive disease is not currently well understood, however it is believed that a number of modes of

action may dominate, such as antibody induced cytotoxicity, cleavage of the extracellular domain of *HER2* and blocking of ligand independent receptor dimerization, among others [130].

Despite the power of Trastuzumab in improving survival for women with *HER2* positive breast cancers, a subset of such women do appear resistant to the treatment [127, 101]. Many hypotheses have been put forward as to why resistance may arise and of these, the possible compensatory signalling via *EGFR* in response to *HER2* inhibition has warranted further investigation [127]. Approximately 30% of *HER2* positive breast cancers are also *EGFR* positive, thus it seems plausible that inhibition of both *HER2* and *EGFR* may prove a more powerful therapy than *HER2* inhibition alone. Consequentially, a dual small molecule inhibitor lapatinib was developed to target both *HER2* and *EGFR* simultaneously [132]. Lapatinib is now FDA approved and has been shown to induce a clinically relevant benefit in patients who are resistant to Trastuzumab [130, 131].

Even with the advances in the treatment of *HER2* positive and HR positive breast cancers, a significant proportion of such cancers remain resistant to therapy. Moreover, there remain between 10 and 20% of breast cancers which show no expression of ER, PR or *HER2*; these tumours are referred to as triple negative breast cancers [133]. These triple negative neoplasms have a very poor prognosis and due to the absence of key therapeutic targets are difficult to treat with efficacy [134]. Consequentially, it is essential to probe the molecular mechanisms of breast cancers in further detail, to attempt to elucidate possible causes of treatment resistance in HR positive and *HER2* positive breast cancers and to identify novel biomarkers and therapeutic targets for the elusive triple negative cancers. The proportions of the different molecular markers found in a large breast cancer data sets are displayed in Fig 1.5B.

1.3.4 Further molecular sub-typing: Genome-wide gene expression

The elucidation of the role of HRs and the *HER2* membrane receptor in breast cancer pathobiology led to great advancements in targeted therapies. However, many patients develop resistance to such interventions and there is still a sizeable subset of individuals for whom no targeted therapy exists. To address this issue many groups have employed high-throughput expression profiling technologies to probe the molecular mechanisms of breast cancer in more detail. Molecular profiling has been conventionally performed at the transcript level by microarrays, though *RNA-sequencing* (RNA-seq) is increasingly being used due to its larger dynamic range and the ability to investigate alternative splicing. Briefly, these assays first require the isolation of RNA (often by spin column techniques

[135]), from a given tumour sample. Following isolation the RNA is reverse transcribed into cDNA. Subsequently the cDNA may be hybridised to a DNA microarray, a chip containing DNA spots in which a large number of sequence specific, fluorescence tagged probes measurably alter their fluorescence when bound to the isolated cDNA, allowing one to quantify the amount of RNA in the sample. Alternatively one may perform RNA-seq, in which the reverse transcribed cDNA is directly quantified by Next Generation Sequencing techniques. Thus these assays give a snapshot of genome wide gene expression in the tumour sample.

By considering the expression of thousands of genes simultaneously, genome wide molecular profiling has led to a more detailed sub-classification of breast cancer than was ever possible by the investigation of a small handful of biomarkers. Their use has provided deeper insights into treatment resistance and differential survival prospects among patients. Moreover, such assays provide the means to investigate which genes drive the worse prognosis subtypes, thus revealing novel targeted treatments .

The first study investigating molecular profiling of breast cancer with the aim of identifying prognostic subtypes was that of Perou *et al.* [136]. In this paper the authors used microarray profiling of 36 IDC-NST and 2 lobular carcinomas to identify 4 molecularly distinct subtypes: luminal, basal-like, *HER2* positive and normal breast like. The authors also suggested a set of 496 genes (involved in specifying endothelial cells, adipose tissue, B and T lymphocytes and macrophages) which displayed intra-subtype homogeneity and yet great variability between subtypes, as a robust classifier of tumour samples into the four groups.

Following this work a larger study was published, which revealed that the luminal breast cancers could be further subdivided into luminal A and luminal B [137]. This discovery pushed molecular subtyping of breast cancer into the limelight of clinical relevance, as it was robustly demonstrated that luminal B breast cancers have significantly worse prognosis than luminal A. In addition to prognostic significance, the subtyping of luminal breast cancer also predicted patient response to treatment. Luminal breast cancers are generally ER positive, and it was demonstrated that the worse prognosis of luminal B breast cancer persisted following tamoxifen treatment, however, it was also found that this subtype responded better to neoadjuvant chemotherapy than luminal A tumours, greatly impacting the treatment of this breast cancer subtype [138]. The proportions of these different molecular subtypes found in a large breast cancer data sets are displayed in Fig 1.5C.

Subsequently, from the consideration of the overlap between human breast carcinoma gene expression and that of mouse models for the disease, the same group of researchers

identified a sixth subtype, similar to basal breast cancer, which was named claudin low [139] and which also had clinical and prognostic significance. As it is of significant relevance to our work on understanding breast cancer at the molecular level, we now explain the hallmarks of each of these subtypes in detail, examining molecular mechanisms and differential responses to therapy.

1.3.4.1 Luminal A Luminal A is the most common subtype of breast cancer, comprising about 35-60% of all cases [138]. Histologically luminal A tumours are generally of low grade, type wise they are reasonably diverse, comprising mostly IDC-NST but also lobular, cribriform, mucinous and tubular carcinomas [118, 111]. This subtype is typically ER and PR positive and *HER2* negative and is also characterised by the expression of genes expressed in the luminal epithelium, as well as *BCL2* and *CK8/18* [138, 137]. Genetically these tumours show a high rate of *PI3K* mutation (49%) and a low rate of *TP53* mutation (12%) [103]. Luminal A tumours also have a low proliferation rate and attempts have been made to characterise this subtype via low expression of the proliferation protein *KI67*, however, this approach has revealed a great deal of inter-observer variability [118].
Luminal A breast cancer has the most favourable prognosis of any breast cancer molecular subtype, with a relapse rate of only 27.8% [138]. This subtype also exhibits a characteristic pattern of metastasis, invading the bone more than any other tissue [138]. Treatment generally involves tamoxifen or aromatase inhibitors (in post-menopausal women), and patients are generally spared chemotherapy, due to the low risk of relapse observed without such treatment [118, 138, 140]. In fact it has been demonstrated that luminal A patients respond less well to chemotherapy than do other breast cancer subtypes [141]. An understanding of this poor response to chemotherapy in luminal A breast cancer has been gained from the consideration of the mitogen activated protein kinase stress response pathway. It was demonstrated that over expression of *MKP-1*, a suppressor of *c-Jun N-terminal kinase* (JNK) signalling, led to chemoresistance in breast cancer cell lines [142]. Subsequently it was shown by next generation sequencing studies, that luminal A cancers have more frequent loss of function mutations in *MAP3K1* or *MAP2K4*, which are essential for JNK signalling [118].

1.3.4.2 Luminal B Luminal B breast cancers comprise between 10 and 25% of all cases [103, 138]. Histologically this subtype is of high grade and of similar type to luminal A. Expression-wise, like luminal A cancers, the luminal B subtype are ER positive, and

express genes associated with the luminal epithelium, however, they tend to be PR negative and may be *HER2* positive as well, they are also often positive for *EGFR* (making them potentially susceptible to treatment with lapatinib) [103, 138]. Luminal B tumours also have a high proliferation rate, and overexpress *KI67* [138]. Genetically these tumours often have mutations in *PIK3CA* (32%) and in *TP53* (29%), in addition they also frequently have *MDM2* amplification (22%) [103, 118]. *MDM2* acts to repress *TP53* and thus its amplification has similar effects to *TP53* loss of function mutation.
Luminal B breast cancer displays a far worse prognosis than luminal A, with a higher probability of relapse and a shorter life expectancy following relapse (1.6 years) [138]. Like luminal A, luminal B tumours are most likely to metastasise to the bone (30%), but they also show a tendency to metastasise to the liver (13.8%) [138]. These cancers also often display resistance to anti-oestrogen therapies despite being ER positive [138] and this resistance can be acquired during the course of treatment. Many strategies to overcome this acquired resistance, such as staggering treatment and dosage and combining multiple forms of therapy have unfortunately proven unsuccessful [118]. Though, as mentioned earlier, luminal B tumours show a better response to chemotherapy than the resistant luminal A tumours (17% *vs.* 7%), this response is still lower than for other breast cancer subtypes (such as *HER2* positive) [138]. An understanding of the molecular mechanisms underlying treatment resistance in luminal B has been gained from considering the molecular profile of this subtype. It was demonstrated by next generation sequencing that inactivation of *TP53* by *MDM2* amplification in luminal B, may contribute to endocrine therapy resistance [143]. Thus inhibition of *MDM2* may prove a powerful option in inducing treatment sensitivity in these cancers [118]. The high rate of mutation in *PI3KCA* has led to further postulation that luminal B tumours may benefit from treatment with inhibitors of the *PI3K/AKT/mTOR* pathway, and such therapy is currently in clinical trials [103].
Lastly, it has been shown that in luminal breast cancers, following surgery, there still exists residual disease in the form of *circulating tumour cells* (CTCs) [118]. Interestingly, it was demonstrated that the self-renewal of these CTCs is dependent on *HER2* expression, even in luminal tumours where *HER2* is not amplified [144]. This suggests a role for anti-*HER2* therapy in preventing disease recurrence following local treatment of luminal breast cancer.

1.3.4.3 *HER2* positive *HER2* positive breast cancers were discussed to some extent earlier and comprise 12-20% of all breast cancers [138]. Though the majority of *HER2*

positive tumours display *HER2* amplification and high *HER2* expression, a significant proportion (30-40%) do not. These tumours are classified as *HER2* positive due to the expression of a similar gene expression profile to that found in *HER2* expressing cancers [103, 130, 138]. This finding is intriguing as it suggests that the *HER2* profile may not be solely driven by *HER2* expression and may be induced by other means, having profound implications for anti-*HER2* therapy and resistance. Histologically *HER2* positive tumours are typically high grade [103]; they are also characterised by poor prognosis and a high proliferative index, though they generally respond well to chemotherapy [103, 130]. *HER2* breast cancers have been further sub-divided into three distinct subtypes, based on gene expression, of which one is of significantly worse prognosis to the others. This poor prognosis subtype is typically ER negative but over-expresses steroid response genes [145]. Though *HER2* positive breast cancer generally responds to trastuzumab, resistant cases exist, moreover, the drug itself is cardiotoxic and thus cannot be administered to certain patients and requires careful cardiac monitoring [130].
Mechanisms of trastuzumab resistance have been investigated and include compensatory activation of the pathways downstream of *HER2* such as the *PI3K/AKT* pathway, either by alternative membrane receptor signalling (*e.g.*, *EGFR*) or by mutation of downstream components [130]. Over 40% of *HER2* positive breast cancers display such mutations in *TP53* or *PIK3CA* [103, 138]. Whilst lapatinib (a dual blocker of *HER2* and *EGFR* mentioned earlier) is able to improve survival in trastuzumab resistant cases where *EGFR* signalling is involved, this drug is poor at treating cancers with *PIK3CA* mutations [130]. Rather intriguingly it was demonstrated that over-exposure to lapatinib may also lead to treatment resistance by the upregulation of *FOXO3*, which in turn upregulates ER [146]. This finding has interesting implications, suggesting that we may supress treatment resistance to lapatinib by combining it with anti-oestrogen therapies.

1.3.4.4 Basal-like Basal-like breast cancer comprises between 10 and 20% of all breast cancer cases [103]. Though these cancers have recently been described via their gene expression profile, the term basal breast cancer has a long history, being originally defined by Moll *et al.* due to histological appearance and the expression of certain cytokeratins [147]. Though there is considerable overlap between basal and basal-like breast cancer it is important to note that they are distinct classifications [148, 149].
Histologically, Basal-like tumours tend to be of high grade and often display lymphocytic infiltrates [149]. These cancers also have a higher tendency to be of metaplastic or medullary histological type [149]. In fact, these breast cancers morphologically resemble

human papilloma virus (HPV) related squamous cell carcinomas. This comparison led to the identification of *RB* under-expression and *CDKN2A* over-expression in basal-like breast cancer, a combination which is associated with *TP53* inactivation [150]. Basal-like breast cancer is named after its tendency to express high molecular weight cytokeratins *CK5* and *CK17*, which are typically expressed in the basal myoepithelial cells of the breast [138], though it has to be emphasised that this is not indicative of cell of origin [148]. Basal-like breast cancers also typically express *EGFR* and genes involved in cell proliferation. Genetically *TP53* and *BRCA1* mutations are also common in this breast cancer subtype [149, 151].

The high rate of *BRCA1* mutations leads to an inherited component to this breast cancer subtype, making it common among African American and Ashkenazi Jewish women. In such patients the cancer presents at a young age and typically at a larger size, with a higher chance of lymph node involvement than other breast cancer subtypes [138, 152]. *BRCA1* is a DNA repair enzyme and mutations in this gene result in a high degree of genomic instability [149]. Moreover, it has been shown that *BRCA1* down-regulation can lead to ER silencing, hence the majority of *BRCA1* mutated breast cancers are ER negative and thus not treatable with endocrine therapy [153]. In fact, the majority of basal-like cancers are triple negative and consequentially there are few targeted therapies making prognosis very poor [134, 103].

Despite the aggressive nature of *BRCA1* mutated, basal-like breast cancer, the knowledge of the role of this gene has led to the development of novel therapies. The poor repair ability of *BRCA1* deficient tumours has led to the use of platinum salts as DNA damage inducing agents, these are cytotoxic to the tumour cells, which cannot tolerate the high level of DNA damage [138]. Moreover, it is known that inhibition of *PARP* genes lead to an increase in the number of single stranded DNA breaks, in the absence of *BRCA1* this damage again induces cytotoxicity [154]. This synthetic lethality has led to the use of *PARP* inhibitors as successful treatments for *BRCA1* mutated breast cancer [153].

1.3.4.5 Claudin-low The claudin-low breast cancer subtype is the most recently identified, hence not much is known about these tumours which comprise 12-14% of all breast cancer cases [138]. Histologically, these cancers are typically high grade IDC-NST cancers, although medullary and metaplastic features are significantly more common than other breast cancer subtypes [64]. Claudin-low breast cancers are characterised by a low expression of tight junction and cell adhesion genes such as E-cadherin and claudin 3, 4 and 7. These tumours are also enriched for *epithelial to mesenchymal transition* (EMT)

genes and immune system response [64].
Claudin-low breast cancers tend to be triple negative (about 80%) and express low levels of proliferation genes [138]. Despite the low proliferation rate claudin-low tumours are of very poor prognosis [64], this is potentially due to the high similarity between claudin-low breast cancers and CSCs [155]. Indeed it was shown by comparison between the expression profiles of mammary stem cells, luminal progenitors and mature luminal cells, that the claudin-low subtype is the least differentiated of all breast cancer subtypes [64]. The possible tumour initiating capacity of claudin-low breast cancer could lead to residual disease following local and adjuvant therapy, re-establishing the tumour at a rapid pace. Therapy for claudin-low breast cancer is currently limited to chemotherapy, though enrichment of *BRCA1* inactivation in this subtype may speak to the utility of *PARP* inhibitors or platinum salts [156].

1.3.4.6 Deeper molecular subtyping Despite the careful characterisation of the above described molecular subtypes, it has become recently apparent that even this level of subdivision may be insufficient. Criticism exists questioning the validity of these classifications, notably from the histological subtype community, where the above described subdivisions are considered only applicable to IDC-NST [108]. Moreover, a large study by the *Molecular Taxonomy of Breast Cancer International Consortium* (METABRIC) profiled genomic aberrations (*single nucleotide polymorphisms* (SNPs) and *copy number variations* (CNVs)) alongside microarray expression data, for 2000 breast cancer tumour samples and concluded that there may in fact be 10 distinct molecular subtypes of breast cancer [157].

1.3.4.7 Gene expression profiling for prognostic evaluation It is clear that molecular subtyping of breast cancer based on gene expression has significant prognostic implications and important clinical relevance in the assignment of an optimal therapeutic regime. Consequentially molecular profiling as a diagnostic tool has made its way into the clinic. Two platforms for assessing gene expression from tumour biopsies to assign a molecular subtype are currently in clinical trials: MammaPrint [158] (FDA approved and mostly utilised in the USA) and OncotypeDX [159] (recently approved by NICE for use in the UK). Both platforms assess the expression of a large number of genes to classify breast cancers into prognostic subtypes and inform treatment. MammaPrint uses a microarray platform to assess a 70 gene signature, mostly of genes involved in proliferation and has proven a robust predictor of distant metastasis and overall survival [103]. OncotypeDX

uses the more precise (but less large scale) *quantitative reverse transcription polymerase chain reaction* (qPCR) assay to assess the expression of 21 genes (5 controls and 16 breast cancer associated, including ER, PR and *HER2*) to provide a similarly robust classification [103].
Though powerful, these gene expression based prognostic assays are limited, as they ignore the important contribution of CSCs in prognostic evaluation.

1.3.5 Oncogenesis and cancer stem cells

We have seen above that breast cancer is highly heterogeneous, with various subtypes characterised by molecular and histological presentation, giving rise to a great diversity in prognosis and optimal treatment regime. Understanding the origins of this heterogeneity requires consideration of how the tumour forms and develops and may provide insights into the prognostic assessment of patients and the investigation of molecular mechanisms for the proposition of targeted therapies.

1.3.5.1 Tumour origins It is typically accepted that a tumour derives from a single cell of origin. This cell is believed to acquire a sequence of successive mutations, eventually culminating in transformation, and subsequently forms the tumour. Several cancers are characterised by the history of this cell of origin and retain markers of its past life, either in histological presentation or in gene expression. Due to the fact that the origin cell must acquire numerous mutations, each in itself a rare event, it is unlikely (but not impossible) for a cancer cell to arise in the number of divisions required for a differentiated cell to stop dividing [160]. It is much more likely, therefore, that an oncogenic cell arises in a cell type that undergoes many divisions during the life of the organism, making the stem cell a strong candidate [161]. This theory has certainly proved valid for some malignancies. One such example is teratocarcinomas, which often contain a large number of highly diverse differentiated cell types and were shown to originate from germ line cells. In addition certain leukaemias are capable of generating every cell in the hematopoietic system, and were found to originate from hematopoietic stem cells [160].
The diversity and heterogeneity of breast cancer, in histological presentation, gene expression and in pathway mutation, is more in line with the haematological malignancies than with other common epithelial cancers (such as colon and pancreatic cancers), hinting at a more multipotent stem cell originator for breast tumours [162]. Consequentially, considerable work has been done to identify possible cells of origin for breast cancers. This search, though hampered by the fact that the differentiation hierarchy of human breast tissue is poorly elucidated, has revealed certain new insights, such as the existence

of a luminal progenitor cell which expresses the ER and may be a suitable cell of origin for ER positive breast cancer [163].
Thus it is likely that healthy stem cells are the source of many malignancies, motivating the work in this thesis on the application of network theoretic tools for measuring cell differentiation potential to oncology. Once a cell of origin has formed, it must next develop into a tumour, understanding of this process is critical to the development of therapies aimed at reversing it.

1.3.5.2 Tumour development There are two key hypotheses regarding tumour development: clonal evolution and the CSC hypothesis [63, 164] (Fig 1.6).

Clonal evolution is a model of cancer development in which tumour heterogeneity is achieved through successive genetic mutation and microenvironmental stimuli. The tumorigenic potential of each cell is thus determined solely by environment and by whether the given cell has acquired permissive mutations in proliferation genes [63]. In the treatment of such cancers, the eradication of every cell must be the goal.
The CSC hypothesis postulates that tumours are organised hierarchically, with a small subset of CSCs endowed with tumorigenic potential whilst the remaining tumour bulk is non-tumorigenic, having differentiated from the CSCs [63]. Under this hypothesis, tumour heterogeneity is a consequence of epigenetic rather than genetic differences among cells [160]. As the cancer cells differentiate they are posited lose their tumorigenic potential, analogous to stem cells during healthy development. Consequentially, therapy under the CSC hypothesis should focus on targeting the small population of CSCs capable of reseeding the tumour rather than the tumour as a whole [165]. It is thus important to consider the defining features of CSCs. A CSC is typically characterised by three criteria: the ability to form tumours in immunocompromised mice, the ability to self-renew and therefore form tumours in secondary mice and the ability to differentiate into cells with non-stem cell characteristics [166].
It is likely that both clonal evolution and CSCs play some role in tumour development and establishing heterogeneity within the tumour, with clonal evolution possibly taking place in the long living CSCs [164]. As explored in earlier sections of this introduction, we wish to develop a network theoretic quantifier of cell differentiation potential, which can provide insights in a cancerous setting. It is clear that a significant point of overlap between the dynamics of healthy development and of oncogenesis lies in the CSC hypothesis. We thus here examine the validity of this postulate in breast cancer in detail.

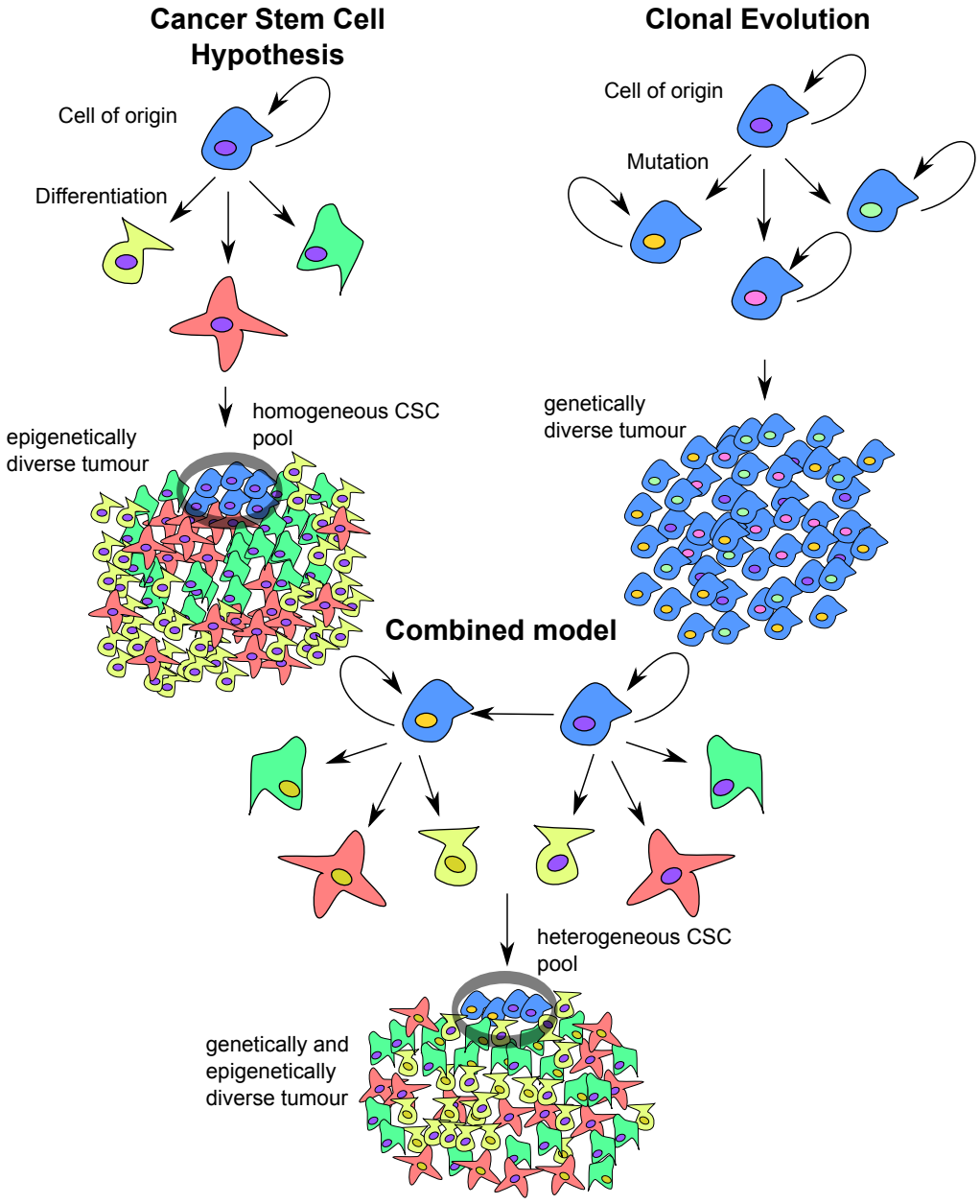

Figure 1.6: **The CSC hypothesis and clonal evolution.** Under the CSC hypothesis, a small CSC pool is capable of self renewal, tumour growth and diversity is driven by the differentiation of the CSC pool into epigenetically distinct cell types, which cannot self renew. In this model, intra-tumour heterogeneity is primarily epigeneitic. Under the clonal evolution model, all tumour cells are capable of self renewal and intra-tumour heterogeneity is driven by mutation. The two models can be combined by considering clonal evolution occurring within the small CSC pool.

1.3.5.3 Evidence for breast CSCs Recently, strong evidence has begun to emerge that breast cancer may adhere to the CSC hypothesis [167]. It was demonstrated by Al-Hajj *et al.* that a small population of breast tumour cells displaying high surface expression of *CD44* and low/no expression of *CD24* (*CD44+/CD24-*) were capable of generating a tumour in immunocompromised mice. In contrast very large populations of breast tumour cells not displaying these markers were incapable of reconstituting a tumour [167]. Interestingly, no morphological phenotype was noted in the breast CSCs, implying that the epigenetic modifications as the tumour differentiated are likely very subtle [164]. Elegantly the CSC population was found to be very rare in the tumour, indicative of a true hierarchical organisation and not simply the loss of proliferation capacity by a small subset of tumour cells via clonal evolution [164].

Breast CSCs were subsequently highly studied and found to display treatment resistance to chemotherapy, hormone therapy and radiotherapy [155, 165, 166, 168]. This intriguing discovery suggested the worrying concept that current treatments were focusing on the wrong cells, attacking the well differentiated, non-tumorigenic tumour bulk and ignoring the dangerous CSCs, which were capable of reseeding the tumour following treatment. A push therefore began to understand the mechanisms of treatment resistance in breast CSCs and many pathways are now being investigated [166]. Of these pathways, those involved in EMT have shown promise in conveying the therapeutic resistance of CSCs. For example, it was demonstrated that increased expression of certain EMT transcription factors (*e.g. Snail1* in breast cancer cells endowed resistance to various chemotherapeutics [165].

Despite this evidence, there are still issues regarding the CSC hypothesis in breast cancer. Notable concerns refer to the immunocompromised mouse assay utilised to demonstrate the tumour initiating capacity of transplanted cells. The mouse generally used is the NOD/SCID mouse [167], which as a result of being immunocompromised is unable to provide the microenvironment of cytokines and immune cells conventionally utilised by a growing tumour. Thus it is difficult to infer whether the cells identified as non-tumorigenic would truly be classed as such were a microenvironment present [164]. It should also be noted that these mice while lacking T cells and B cells do retain natural killer cells, which can reject transplanted human cells and thus under-estimate the proportion of CSCs [169]. Thus more rigorous assays are required to identify whether the tumorigenic population identified is indeed a minority of the tumour, indicative of hierarchical organisation. In addition, given the diversity of breast cancer it is reasonable to ask: do CSCs in all breast cancer subtypes have the same *CD44+/CD24-* signature? Several studies have suggested that this is likely the case for basal breast cancer and possibly *HER2* positive

breast cancer, but is unlikely for ER positive breast cancer [170]. This could be due to different cells of origin for these distinct breast cancer subtypes and raises the need for less biomarker specific, more general systems level markers of CSCs. A more conceptual concern is the recent notion of cancer cell plasticity [63], in which non-CSCs can switch to become CSCs under certain conditions. Evidence for such transitions in breast cancer have been found, and it was recently demonstrated that the chromatin state of the *ZEB1* promoter can control this plasticity [171]. Cancer cell plasticity undermines the treatment strategy of the CSC hypothesis, *i.e.*, targeting the small number of tumorigenic cells to eliminate the tumour. If it is possible for a non-CSC to become a CSC, then all of the tumour must be considered tumorigenic, unless the mechanism underlying plasticity can be understood and targeted [63].

Thus it is clear that healthy development, via stem cells, likely plays an important role in the birth and growth of a tumour. It is thus logical that a network theoretic tool developed to understand cell differentiation will also generate important insights into the molecular mechanisms of oncogensis.

1.3.6 Perspectives: Relevance for our research

The aim of our work in reference to breast cancer is to develop and apply methodologies which might postulate novel therapeutics. In this review we have examined the highly heterogeneous nature of breast cancer pathology and how optimal treatment strategy is critically dependent upon tumour subtype. Consequentially the application of any methodology aimed at elucidating useful therapeutics must consider these highly diverse subtypes separately.

We saw in the first review on network methodologies, how network theory has been applied to cancer, revealing the importance of hub gene mutations in initiating transformation and revealing pathway dysregulation. However, we also saw that very few methodologies focus on understanding general principles surrounding the network rewiring in cancer progression.

We believe that cell differentiation is a very important process in cancer, which may be amenable to investigation via network theoretic tools. In the penultimate section of this review we explored the CSC hypothesis of tumour organisation and discussed the evidence for its validity in breast cancer. We saw how breast CSCs are resistant to conventional therapeutics, and thus appear important candidates for investigation by methodologies aimed at identifying new therapies. We also explored the emerging concept of cancer cell plasticity and saw that it is possible for certain epigenetic modifications to convert non-tumorigenic cells to CSCs.

It seems likely that as in normal development, the differentiation of CSCs into non-tumorigenic cells and indeed the reversion of this process by plasticity, involves some global coordinated biological network rewiring. Characterisation of this rewiring may be achieved by the consideration of cell potency in healthy development. As we explored in the second introductory section on the mathematics of stem cell biology, network entropy may prove a powerful tool in the quantification of cell potency and the development of measures to compute drivers of cell differentiation. If tumour development indeed represents a caricature of healthy development, one would expect network entropy to prove similarly powerful in discriminating the critical drivers of CSC differentiation and cancer cell plasticity. These drivers could represent important drug targets, capable of preventing tumours replenishing their stem cell pool by plasticity and simultaneously blocking tumour differentiation. Moreover network entropy may provide an unbiased classifier for CSCs, not reliant on the expression of certain markers selected on an ad hoc basis.
In addition to approaches centred around the CSC hypothesis, the analysis of tumour cell potency may prove informative in understanding pathway activation differences and treatment responses between the various phenotypes of breast cancer. Breast carcinomas display a wide diversity in their level of differentiation. Currently this level is assessed by histological grade as discussed above which is a strong prognostic indicator. This qualitative scoring based on a number of morphological features, however, is both laborious to measure and sheds no light on the molecular mechanisms driving the associated disease severity. If entropic measures are capable of accurately recapitulating tumour grade, however, they provide an unbiased assessment of tumour differentiation and have the potential to identify molecular drivers of tumour stemness, permitting one to posit therapeutics for arresting tumour development. Consequentially, we will focus on the application of entropy based network rewiring methodologies to human breast cancer and related malignancies.

1.4 The pathophysiology of FSHD

In this thesis we aim to utilise network theoretic tools to investigate the pathomechanisms of complex disorders involving differentiation. In the previous introductory subsection we considered human breast cancer, a malicious caricature of healthy development, in which the power of cellular differentiation is usurped and utilised to the detriment of the organism. The task of the cancer investigator is to find therapies which can inhibit the development of pathological tissue. On the other side of the pathology coin are disorders in which healthy differentiation is impaired, and the role of the investigator is to find therapies to restore the normal development of healthy tissue. Of such disorders, arguably

the most well known are the muscular dystrophies.

The muscular dystrophies are group of inherited myopathies, which show considerable heterogeneity in clinical presentation [172]. Though different muscular dystrophies affect distinct muscle groups, with varying severity, all are characterised by a progressive skeletal muscle atrophy, resulting in a state of continual skeletal muscle breakdown. Skeletal muscle comprises of many bundles of multi-nucleated myofibres, which are heterogeneous in type, permitting a wide array of movements [173]. Myofibres contain sarcomeres which are the functional units of skeletal muscle, comprising of two classes of myofibrillar proteins: myosins and actins. The interactions between actins and myosins provide the force required for muscle contraction. Each individual myofibre is contained within a thin layer of connective tissue, referred to as the endomysium, whilst myofibre bundles (fascicles) are bound within another connective tissue layer, the perimysium. An entire skeletal muscle in bound by a final layer of connective tissue the epimysium, which maintains the shape of the muscle. These connective tissue layers fuse at the muscle terminus, forming tendons which typically anchor the muscle to bone (Fig 1.7).

The state of continual skeletal muscle breakdown in muscular dystrophy demands constant repair of myofibers by differentiation, a process known as myogenesis. The main skeletal muscle stem cell is the muscle satellite cell, which occupies a niche beneath the basal lamina [174] (Fig 1.7). Quiescent muscle satellite cells express the transcription factor *PAX7*, however, upon activation by muscle damage, the cells migrate from the basal lamina and undergo asymmetric division, with a proportion of cells expressing the basic helix loop helix, myogenic regulatory factor *MYOD*. The cells expressing *MYOD* (known as myoblasts) proliferate and lose *PAX7* expression, eventually expressing a later myogenic regulatory factor *Myogenin*, which terminates myoblast proliferation and initiates alignment and fusion into multi-nucleated myofibres, which express myosins, such as myosin heavy chain (*MyHC*). In certain muscular dystropies, continual damage is thought to put strain on the satellite cell pool, whilst in others the entire process of myogenesis is believed to be disrupted. It is to a prominent member of the latter class which we now turn.

Facioscapulohumeral muscular dystrophy (FSHD) is the third most common muscular dystrophy with an incidence of around 1/20000 [175], however, the relative longevity of patients ensures that this condition is in fact the most prevalent form of muscular dystrophy, with a prevalence of 4-12/100000 [176]. Despite such prevalence no treatment currently exists for this complex and multifaceted disorder, (though several have been suggested [177, 178, 179, 180, 181]). One of the current greatest barriers to the discovery of a viable treatment is the elucidation of the enigmatic genotype to phenotype link.

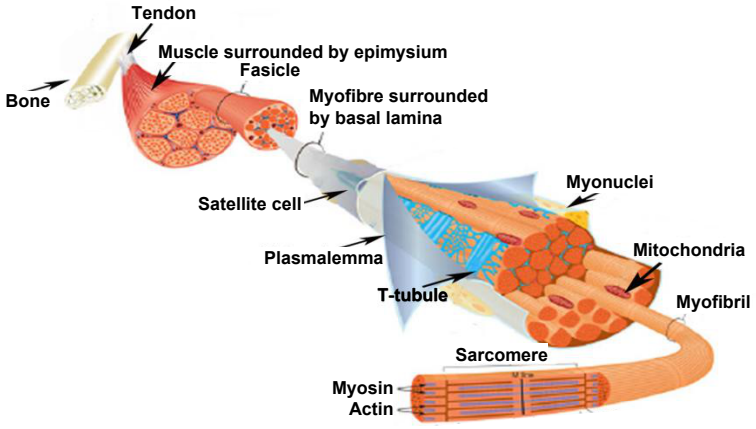

Figure 1.7: **The structure of skeletal muscle.** Figure adapted from [174].

FSHD is an autosomal dominant inheritable disorder typically diagnosed by the second decade of life. The condition is progressive and approximately 20% of patients eventually require the use of a wheelchair [182]. The condition manifests most notably via an asymmetric, highly tissue specific skeletal muscle atrophy, initially effecting the facial muscles, then extending to the limb girdle and upper arm muscles [183] with occasional lower limb involvement [184]. However, FSHD is also associated with a number of extra-muscular features, pointing to a more global dysregulation of signalling pathways in the pathology. The best documented of these unusual features are a retinal vasculopathy very similar in presentation to Coat's disease [185, 186, 187], and a high frequency, sensorineural hearing loss, believed to be cochlear [188, 189, 190], both of which have been associated with early onset or severe FSHD. Perhaps less well studied are subclinical cardiac arrhythmias, which have also been found to associated with the pathology [191]. It was noted that any unifying hypothesis of FSHD pathomechanisms must explain both the characteristic muscle wastage and these less common symptoms, and must therefore consider a systems perspective [186]. It is clear, therefore that FSHD is a prime candidate pathology for the application of network theoretic tools.

In understanding a complex developmental pathology, it is critical to first consider the clinical presentation in detail. This phenotypic manifestation may provide critical clues to pathway dysregulation. Further, when considering the molecular mechanisms of an inheritable condition, insights from genetics often prove invaluable. In what follows, therefore, we first explore the history of FSHD research, from its clinical presentation, through to the identification of its genetic basis. Subsequently we explore the primary candidate FSHD gene *DUX4*, elucidated from the genetic basis of FSHD, and examine endeavours

to understand how very low expression of this gene may cause the pathology. We then examine the results of a more general probing of FSHD molecular mechanisms via muscle biopsy genome wide gene expression. We close as usual with perspectives relating the findings of this literature review to the work undertaken in this thesis. In particular we note that network based approaches have been distinctly absent in FSHD research and we argue that they are essential to unravelling the molecular mechanisms of this highly systemic and heterogeneous pathology. Moreover, we note that an entropic approach may determine the impact of *DUX4* expression on muscle differentiation dynamics.

1.4.1 History of clinical presentation and heredity

Early understanding of muscle disease was hampered by the belief up until the mid 1800s that the basis of all pathological muscle wastage was neurological. Thus, as a myopathy, with no affectation of the spinal muscles, nor signs of myelopathy in biopsied samples, FSHD was not identified until the latter half of the 19^{th} century. The first clinical report of a patient with FSHD is believed to be due to Cruveilheirs in 1848 and describes an 18 year old male, who died of smallpox, and presented with severe weakening of the facioscapulohumeral regions, with no spinal involvement [184]. Though this report is likely the first description of the pathology, it received little attention at the time of publication and it was not until 36 years later in 1884 that Landouzy and Dejerine, published the first complete description of FSHD in a single family [192]. Landouzy-Dejerine disease, as FSHD was initially known, was described in the original family as a myopathy without nervous system involvement that began in the face and extended over life to the upper limbs. Over time more detailed descriptions of patients arose and a more robust picture of clinical pathology was gained. Facial weakness was generally observed earliest, with a notable weakening in the orbicularis oris and orbicularis oculi manifesting in an inability to fully close the eyes, and a weakness in the muscles around the mouth often resulting in the inability to whistle [184]. Subsequent upper limb involvement derives from affectation of the latissimus dorsi, serratus anterior, pectoralis, subscapularis, rhomboids and latterly the trapezius [184, 193]. Lower limb involvement typically presents as foot drop and involvement of the pelvic girdle muscles [184], the rare and heterogeneous involvement of the lower limbs led some physicians to suggest a separate classification from FSHD [194]. It should be emphasised that all muscle atrophy in FSHD often presents asymmetrically, without a consensus severity bias for left or right [184] (Fig 1.8).

In addition to muscle involvement, it has been demonstrated that FSHD is significantly associated with an exudative retinal vasculopathy similar in presentation to both Coat's

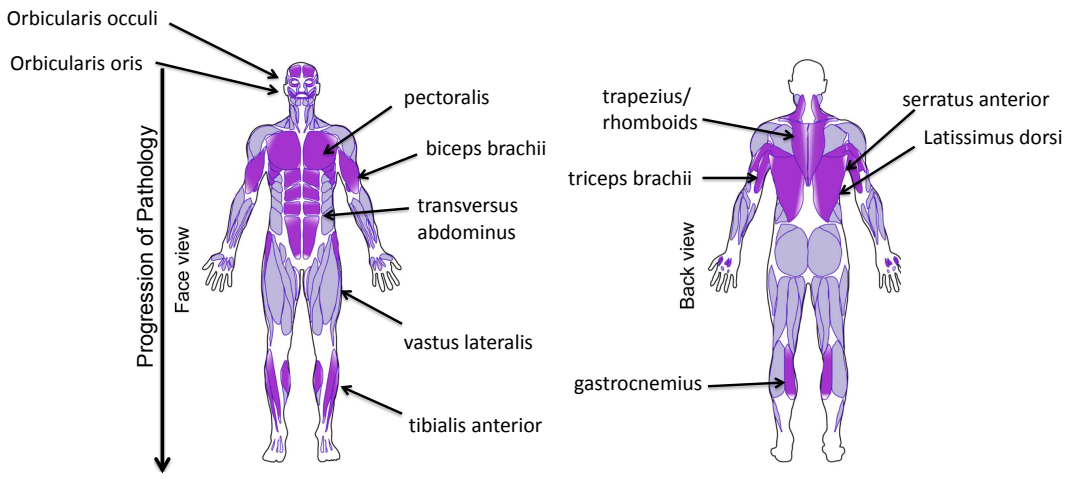

Figure 1.8: **FSHD affected muscle groups.** This figure was produced by Francoise Helmbacher and was adapted with kind permission.

disease and Norrie Disease [185, 186, 187, 195, 196]. Coat's disease only effects males and typically presents unilaterally as a retinal telangiectasia consequential of a breakdown of the endothelial cell wall of blood vessels in the retina, though asymmetric cases have been reported [197]. It is progressive and can lead to blindness through retinal detachment [186]. Norrie Disease is similarly X linked and presents as a failure of vascularisation of the retinal periphery, alongside telangiectatic blood vessels and again can lead to blindness via retinal detachment [186]. Interestingly, Norrie disease is also associated with an asymmetric, high frequency, sensorineural hearing loss, believed to be cochlear on the basis of mouse models [198]. Whilst about half of FSHD patients display some sort of retinal vasculature abnormality as detected by florescence angiography [196], these tend to be subclinical in the vast majority of patients. Indeed, a recent meta-analysis of Coat's disease in FSHD demonstrated that it was a rare symptom associated with severe presentations of FSHD [196]. Moreover, in 70 reported cases of Coat's Disease in FSHD, only a single patient reported visual loss [196]. In addition, retinal symptoms in FSHD can be bilateral and equally effect women and men, suggesting a deviation from Coat's disease and Norrie Disease [199]. Thus whilst some similarities exist between the clinical presentation of FSHD and these retinal conditions, suggesting an overlap in their molecular mechanisms, their presentation is clearly less severe in FSHD.

Sensorineural hearing loss is also associated with severe cases of FSHD which present early in childhood [189], however, there is controversial literature on the association with

typical onset FSHD [188, 190]. When observed, hearing loss tends to be bilateral and progressive, in a manner similar to noise induced hearing loss, though asymmetric presentations have been documented [196]. As in Norrie disease, hearing loss in FSHD is believed to be cochlear [186].

There have also been reports of cardiac abnormalities in FSHD, including presentations of hypertrophic cardiomyopathy [200]. The only significant cardiac defect associated with FSHD, however, appears to be an arrhythmic alteration [191]. This alteration is often reported as subclinical, though it has been suggested to be progressive with FSHD in a manner that may increase the risk of a cardiac event [201, 202].

In addition to a report of clinical presentation, Landouzy and Dejerine, the authors of the initial study on FSHD, also noted that the condition appeared to show an autosomal dominant inheritance pattern in their family of study. This observation was subsequently confirmed in 1950 in a study of 4 families exhibiting 18 cases of FSHD [193], and again in 1982 by a study of a further 19 families by Padberg [183], in which it was concluded that a recessive pattern of inheritance was of negligible likelihood.

The next challenge of FSHD research became the hunt for the susceptibility locus, however, this was not as straightforward as expected. Initial studies by Padburg on 120 individuals from 10 families using early genetic linkage techniques failed to identify any significant loci. Following the development of markers for microsatellite regions is was eventually demonstrated that FSHD was significantly associated with a locus at 4q35 a sub-telomeric region of chromosome 4q [203]. It was subsequently demonstrated in 1993 that FSHD was associated with a contraction of a macrosatellite repeat sequence termed D4Z4 at the 4q35 locus [204]. Between 1 and 10 D4Z4 repeats were required to convey susceptibility to FSHD, with healthy individuals having up to 100. Interestingly, although repeat length was negatively correlated with disease severity [205], at least one repeat appeared essential for FSHD, with complete loss of the D4Z4 region not associated with the pathology [206]. Critically, it was also shown that each D4Z4 repeat encoded two homeoboxes, which displayed the same sequence in pathological and control individuals and early studies were unable to report any expression of the region in either healthy or FSHD samples [204]. The fact that FSHD was a disease of copy number variation rather than a mutation in a functional gene represented a great barrier to the understanding of the pathomechanisms of the pathology. The D4Z4 region was considered junk DNA with no role in healthy function, thus contractions in this region could not be effecting normal development in a trivial manner. Moreover, there was a small subset of FSHD patients (about 5%), who presented with the clinical phenotype but without D4Z4 contraction, indicating multiple mechanisms of pathogenesis [207]; this subgroup is generally referred

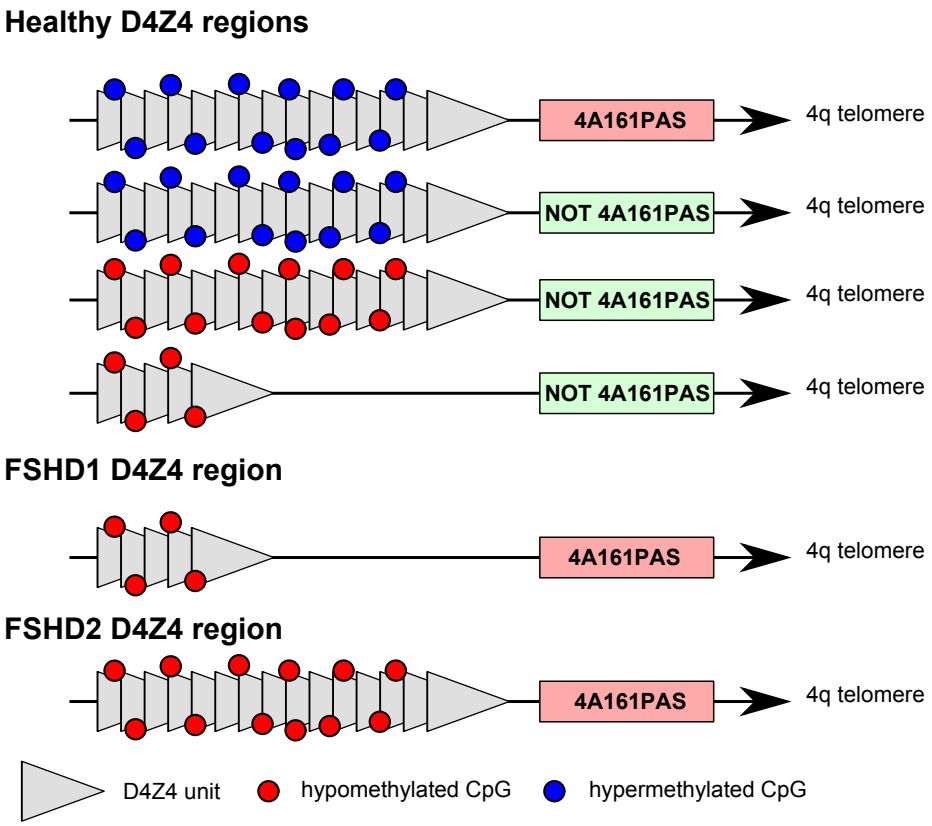

Figure 1.9: **The Genetics of FSHD.** FSHD arises via hypomethylation of the D4Z4 region on chromosome 4q35, given a permissive haplotype (4A161PAS) encoding a polyadenylation signal. Hypomethylation can occur either via truncation of the D4Z4 region (FSHD1), or via other mechanisms, such as mutation in *SCHMD1* (FSHD2).

to as FSHD2 (OMIM158901), with the D4Z4 contraction associated group referred to as FSHD1 (OMIM158900). FSHD2 was associated with hypomethylation of the D4Z4 region implying that mis-regulation of this region may be the basis of the clinical presentation of FSHD [208, 209, 210] (Fig 1.9). However, as we will see, it was many years until sufficiently large FSHD2 patient cohorts permitted the identification of the genetic basis of this rare subtype. The complex genetic nature of FSHD represented an ominous stumbling block to the identification of candidate genes to target to mitigate the pathology.

1.4.2 The identification of FSHD candidate genes and the rise of *DUX4*

For many years it was believed that the D4Z4 repeat region was non-coding and the general hypothesis of FSHD1 pathogenesis was that D4Z4 truncation conveyed pathology through a chromatin destabilisation, leading to the aberrant expression of chromosome 4q35 neighbouring genes, such as FSHD region gene 1 (*FRG1*) [211, 212]. Investigations along these lines were initially promising, and it was shown that mice over expressing *FRG1* displayed a muscle phenotype [211]. Expression results were not always consistent with this hypothesis of D4Z4 regional dysregulation, however, and it was called into doubt strongly in an early microarray study [213]. It was discovered in 1999 that each 3.3kb D4Z4 repeat encoded a copy of a novel transcription factor, double homeobox 4 (*DUX4*) [214]. Structurally *DUX4*'s N-terminus contains two homeodomains with high similarity to the homeodomains of the transcription factors *PAX7* and *PAX3*, (important in muscle satellite cell specification), and the C-terminus is an activator of transcription [175, 215]. *DUX4* admits two stable (polyadenylated) splice variants, a full length transcript, *DUX4*, a shorter transcript, limited only to the homeodomains, *DUX4s*, in addition, there exists a homologous sequence mapping 42kb centromeric of the D4Z4 repeats termed *DUX4c*. Only the full length transcript is differentially expressed in FSHD, with the other two variants expressed in both FSHD and control samples [216]. Interestingly *DUX4c* appears to be induced specifically during FSHD myogenesis, moreover it has been shown to regulate myogenic factors, suggesting a role in pathology [217, 218]. Phylogenetic studies revealed that *DUX4* was evolutionary conserved in primates [219] and a mouse homologue, *mDUX*, has recently been shown to induce a myopathic phenotype when over expressed in mice [220]. However, the 4q35 D4Z4 region is primate specific and there exists no syntenic region in the mouse [221].

These findings initiated a new hypothesis of aberrant expression of *DUX4* leading to broad transcriptional dysregulation culminating in the FSHD phenotype. Subsequent studies investigated targets of *DUX4* and it was demonstrated that pituitary homeobox 1, (*PITX1*) was a direct target [222]. It was later revealed that *PITX1* causes skeletal muscle dystrophy in mice when over-expressed [222, 223]. Issues with the *DUX4* hypothesis of FSHD1 pathogenesis also arose from the fact that *DUX4* expression was not observed consistently in FSHD1 myoblasts. It was not until 2010 that it was demonstrated by the use of qPCR, with high quantities of cDNA template, that *DUX4* was expressed in FSHD1 myoblasts (albeit in very small quantities), though not in healthy controls [224]. It was also revealed that *DUX4* expression required a permissive haplotype (4A161PAS), which was demonstrated to present in both FSHD1 and FSHD2 patients. The haplotype consists of a polyadenylation signal adjacent and distal to the subtelomeric D4Z4 repeat

region [224]. This chromosomal background is essential to ensure the stability of *DUX4* mRNA transcripts expressed from the most distal D4Z4 unit, a finding that provided strong evidence for *DUX4* as an FSHD1 candidate gene.

Given the potentially important role of *DUX4* in FSHD1, much work was subsequently undertaken to identify the genetics of FSHD2 with the hope of reconciling this condition with FSHD1 via the *DUX4* hypothesis. It was recently demonstrated that FSHD2 was significantly associated with a mutation in *SMCHD1* a putative epigenetic modifier, which was shown to induce D4Z4 contraction independent *DUX4* expression [225]. This finding provided strong evidence for the involvement of *DUX4* in both FSHD subtypes, reconciling the similar clinical phenotypes with disparate genetics.

Many studies began to investigate the mechanisms by which *DUX4* may induce FSHD, generally by cell culture experiments, documenting the *DUX4* over-expression phenotype. These revealed that *DUX4* was highly toxic in murine myoblasts, and that this toxicity was p53 dependant [181]. Oxidative stress sensitivity was also shown to be induced by *DUX4* by the generation of a doxycycline inducible *DUX4* expressing C2C12 myoblast cell line [175]. This study further demonstrated that *DUX4* expression caused oxidative stress sensitivity (at least in part) by repression of the glutathione redox pathway. *DUX4* expression also induced a myogenesis defect and this was shown to be related to the over expression of MyoD [175]. Moreover, *DUX4* expression in ES cells was found to push cells down the neuronal lineage [226].

Whilst encouraging a complication to the understanding of FSHD pathogenesis through the study of *DUX4* centres around the fact that the gene is only expressed at very low levels in FSHD muscle [227]. Moreover, it was demonstrated that this low abundance is attributable to high expression of *DUX4* by a small proportion (as low as 1/10000) of FSHD myoblasts [216]. It is therefore natural to ask how this low concentration of *DUX4* can lead to the symptoms FSHD. Two theories currently stand on this issue: the minority rules hypothesis and the majority rules hypothesis [228]. In the minority rules hypothesis FSHD is driven by the high expression of *DUX4* in a small number of myoblasts and myotubes. *DUX4* is posited to cause such a widespread phenotype by diffusing out of single nuclei into the cytoplasm of multinucleated myofibers, where it can enter adjacent nuclei and spread its transcriptional dysregulation throughout the fibre. Evidence for this hypothesis stems from the clustering of *DUX4* positive nuclei in cultured FSHD myoblasts. Moreover, it was recently demonstrated that *DUX4* expression in a single nucleus of a multinucleated myotube leads to a gradient of expression of *DUX4* and PITX1 in neighbouring nuclei [227].

Conversely, the majority rules hypothesis posits that transient high expression of *DUX4*

occurring in all cells in the myogeneic lineage before the myoblast stage leads to FSHD, and that this expression is inefficiently silenced in a small proportion of cells [228]. The majority rules hypothesis draws credit from the short half life of *DUX4*, regulated by the ubiquitin proteasome pathway [228], which makes long range diffusion of *DUX4* unlikely. Moreover, the involvement of *DUX4* in myogenesis, would lead one to believe that its pathogenic effects are upstream of the myofibre stage, at which point the minority rules hypothesis is invalid.

Investigation of the role of *DUX4* in muscle differentiation may prove critical in informing which of these postulates are correct, and hence providing a stronger basis for the development of therapies. A transcriptomic measure of cell potency, similar to that explored in previous sections, may be able to determine whether *DUX4* over expression retards or accelerates the differentiation process in FSHD muscle.

1.4.3 *DUX4*, necessary but not sufficient? An emerging role for telomeres

We have above presented evidence for the critical role of *DUX4* in FSHD pathogenesis. Recently evidence has begun to accumulate, however, demonstrating that though *DUX4* may be necessary to cause FSHD, other factors may also be at play. Evidence for this theory came from a recent study of Italian families, in which the truncated D4Z4 regions with the permissive FSHD allele were found at a prevalence of 1.2%, one hundred times higher than the prevalence of FSHD [229]. In addition, *DUX4* expression was found in the muscle cells of both FSHD patients and their unaffected first degree relatives, indicative of *DUX4* modifiers determining disease progression [230].

A new hypothesis has emerged to reconcile these observations, centred around the concept of telomere position effect (TPE) [231]. TPE is an epigenetic silencing mechanism in eukaryotes, in which repressive heterochromatin spreads from the telomeres over the sub-telomeric region causing gene silencing, as the telomere shortens, this repressive effect diminishes as the heterochromatin shrinks [232]. As *DUX4* is located sub-telomerically on chromosome 4q it was postulated that telomere length variability and truncation, with numerous cell divisions, may contribute to the heterogeneous nature (even among monozygotic twins) and progression of FSHD [231]. In fact it was demonstrated that telomere length was significantly negatively correlated with *DUX4* expression in immortalised FSHD myoblasts, marking FSHD out as the first condition in which TPE has been demonstrated as influential to pathology.

1.4.4 FSHD transcriptional dysregulation

Given that *DUX4* expression may not be the complete story, many investigators have considered expression profiling of FSHD muscle biopsies, to provide information on transcriptional dysregulation unbiased by the *DUX4* hypothesis [52, 187, 213, 233, 234]. In line with *DUX4* cell culture studies, the majority of these gene expression analyses on FSHD muslce biopsies implicated the transcription factor MyoD. Moreover, FSHD myoblasts also demonstrated a differentiation defect with myotubes from FSHD patients exhibiting two distinct phenotypes in different proportions, an atrophic phenotype and a disorganised phenotype [235, 236]. It was also revealed that as with *DUX4* expressing myoblasts, FSHD myoblasts displayed a sensitivity to oxidative stress [235]. Several muscle biopsy gene expression studies have also implicated members of the JNK signalling pathway as significantly differentially expressed in FSHD muscle, such as *JUND* [187], *JUNB* [52] and *JUN* [213]. Immune system pathway dysregulation is also observed in FSHD [187, 234] and two studies [177, 180] have reported the over-expression of TNFα. Genes involved in RNA processing and the ubiquitin pathway have also been reported as differentially expressed in FSHD [234], while several studies have implicated genes involved with the cell cycle and apoptosis particularly via p53 [52, 181, 187]. Sensitivity to oxidative stress is a well-studied phenotype of FSHD muscle cells, and there has been several efforts made to investigate the dysregulation of the hypoxic response in FSHD [180, 235]. Notably the abnormal over-expression of many anti-oxidant enzymes was demonstrated in FSHD (although notably not MnSOD), as well as mitochondrial dysfunction potentially attributable to *cytochrome c oxidase* (COX) activity decrease [52, 180]. In addition the differential expression of a number of genes associated with hypoxia inducible factor 1 α (HIF1α) signalling and the NRF-2 mediated oxidative stress response in FSHD was observed [52].

Given the great diversity of pathway perturbations reported in FSHD, the elucidation of a unifying theory of pathogenesis has remained enigmatic. Most attempts at such a theory have been largely hypothesis driven, rather than data driven and include dysregulation of Wnt/β-catenin signalling [186], Ca^{2+} signalling [177], epigenetic regulation [210], perturbation of the nuclear envelope [213], disruption of p53 activity [181] and RAGE-NF-κB signalling [237]. An understanding of which molecular mechanisms may prove the most therapeutically viable is also currently lacking.

Clearly the dysregulation of such a large number of distinct genes and pathways, yet the positing of a causal aberration arising from mis-expression of a single gene, suggests that there may a systems property of network rewiring that can trace the diverse molecular pheonotype of FSHD from its causal genotype. Clearly differentiation is also perturbed

in FSHD yet it is unclear what targets may be the best placed to understand this. The description of molecular networks which are perturbed to convey FSHD pathology in a manner related to *DUX4* has thus been regularly set as a key priority of FSHD research.

1.4.5 Perspectives: Relevance for our research

The aim of our work in reference to FSHD is to develop and apply methodologies which might postulate novel therapeutics, as well as validate these experimentally. In this section we have examined the highly diverse nature of the clinical presentation of FSHD, indicating that it is a condition affecting a variety of tissues in different qualitative ways. We have also seen how the complexity of this condition both genetically and mechanistically has hampered the development of viable therapeutics.

Though a candidate gene has been identified in the transcription factor *DUX4*, research into the effect of this protein on muscle cells has not yet been approached from a global network theoretic perspective. Moreover, though there is a clear muscle differentiation defect in FSHD, precisely how this defect causes muscle weakness in patients and the role of *DUX4* therein, is poorly understood. The application of an entropic measure of differentiation potential discussed in earlier sections, may prove critical in understanding the effect of *DUX4* overexpression on myogenesis. Whilst careful examination of FSHD myoblast differentiation may provide insights into the FSHD myogenesis defect.

Understanding the role of *DUX4* in FSHD pathology is complicated by the fact that, though present in FSHD muscle, the *DUX4* protein and transcript are only found at very low levels [227]. Such a result presents a problem for biochemical assays as it often makes non-primary cell culture models focused on *DUX4* over-expression somewhat unrealistic. Hence muscle biopsy gene expression data presents a picture of FSHD pathomechanisms unhindered by non-physiological *DUX4* levels. Considerable such data has been produced, revealing a number of disjoint pathways involved in FSHD, however, an integration of these pathways and the role of *DUX4* is lacking. There is thus scope for the application of network theoretic tools, aimed at elucidating a unifying picture of FSHD pathomechansims, and understanding *DUX4*'s role in FSHD pathogenesis.

Though progress has been made towards generating FSHD animal models, the current most viable candidate (a mouse model) lacks a muscle phenotype [238]. Thus without a suitable proxy, the most sensible way to analyse and understand FSHD pathomechanisms is through the consideration of primary muscle biopsies, or from patient derived cell lines. In such a rare condition the acquisition of a sufficient number of primary samples from patients for reliable inference is an unappealing task. Scapular fusion surgery, an intervention to address scapular winging in FSHD involving fixation of the scapular

to the ribcage, is a low risk procedure associated with improved mobility and reduced shoulder pain and fatigue [239, 240]. The precise procedure varies, but always requires the clearing of the subscapularis prior to fixation [178, 239]. The resulting large primary muscle sample is generally disposed of, eliminating a rare and valuable research resource. Moreover, the surgery is routinely performed twice on the same patient, once on either shoulder, thus samples obtained from these procedures represent an elegant opportunity to investigate asymmetry in FSHD. Obtaining such such samples was sadly beyond the scope of this thesis. However, in lieu of these, we believe the best model available for interrogating FSHD myogenesis are the immortalised mosaic myoblast cell lines isolated from a single patient, which are isogenic with the exception of D4Z4 repeat contraction [241]. It is important to note that these cells require more thorough characterisation before use.

Thus our work on FSHD will focus first on the application of entropic network measures to gene expression data induced by *DUX4* over-expression in muscle cell precursors, to ascertain whether this gene increases or decreases the stemness of a myoblast, thus inhibiting or accelerating myogenesis to cause the FSHD phenotype. Subsequently, we will characterise the mosaic cellular model of FSHD, confirming *DUX4* over-expression and examining experientially, whether muscle regeneration is indeed slowed or accelerated as predicted by network theoretic measures. Finally, we will utilise entropy based network theoretic tools to perform a meta-analysis of FSHD muscle biopsy gene expression data, we will aim to construct a unifying model of FSHD pathomechanisms and posit and validate therapeutic targets, as well as comparing our model to the genes perturbed by *DUX4* expression.

1.5 Overview of introductory sections

In this quite extensive introduction we have covered, what on the surface may appear 4 seemingly disjoint topics. However, we hope that by providing this extensive background, the reader can appreciate that our journey through these distinct subjects is actually very natural.

We began with network theory and its application to complex pathology. We revealed that though clear evidence exists that biological networks rewire in some ordered manner in response to internal and external perturbation, little attempt has been made to characterise a systems property of this rewiring and exploit it to better understand network biology, with the possible exception of the under-developed network entropy.

We next moved on to study development and cell differentiation, processes ubiquitous to multi-cellular life and central to many complex pathologies. We saw that experimental

evidence hinted at a systems property that changed systematically during cellular differentiation, the randomness of gene expression. From these first two sections we formed an initial plan: develop the network entropy methodology and validate it as a way to measure a systems property of network rewiring that alters predictably during cellular differentiation.

Given that our primary aim is to apply network theoretic tools to complex disease, and that the most logical choice of network theoretic measure correlates with differentiation potential, it is only natural to consider disorders of development as pathologies of interest. In the final two sections of our introduction, therefore, we explored two ends of the developmental pathology spectrum. Firstly we considered cancer, in which the power of cell differentiation is hijacked, to develop a malicious new tissue. Secondly, we considered muscular dystrophy, in which cell differentiation is inhibited, resulting in the poor development of muscle tissue.

We explored in detail what is known and what is unknown, about a key pathology from each class: breast cancer and FSHD. We indicated where our differentiation based network theoretic tools may prove most useful and outlined approaches to improve understanding of areas that require elucidation.

In what follows we execute the plan of action detailed in this introduction, first developing entropy based network theoretic tools, then validating the most viable candidate as a measure of cellular differentiation and finally applying our methods to understand breast cancer and FSHD pathomechansisms. We hope that we have been able to describe many novel findings and have laid some foundations from which we can advance the field.

2 Entropic Network Theoretic Tools: Concept and Theory

In the introduction we motivated the need to develop entropy based network rewiring methodologies, for improving the understanding of systems properties of biological interactions in developmental disorders. We concluded that a protein interaction network (PIN) compiled from experimental data, but refined to remove interactions between organelle separated proteins, would serve as the most suitable model of signal transduction architecture. We also promoted the use of an edge weighting, derived from genome wide gene expression data as a means to introduce sample and phenotype specificity to our interaction model. Finally, we emphasised the need to develop tools which were scalable, in the sense that they could elucidate global principles of biological network rewiring, whilst also identifying critical genes and pathways.

To this end, in this chapter we first motivate and introduce a simple random walk model of traffic on a data weighted PIN, which can be represented as a stochastic matrix. We then explore 3 distinct entropy based network theoretic tools for the interrogation of biological network rewiring, namely:

1. **Network Transfer Entropy (NTE):** A Jensen-Shannon Divergence based measure for detecting the amount of information transferred between two vertices on a weighted PIN.

2. **Signalling Entropy:** A measure for quantifying global signalling promiscuity in a weighted PIN based on the entropy rate of a random walk process.

3. **Interactome Sparsification and Rewiring (InSpiRe):** A network rewiring algorithm which utilises vertex specific entropy and Kullback-Leibler divergence to detect network rewiring between biological phenotypes described by a weighted PIN.

Following introduction of the weighted random walk model of network traffic, we first explain the development of NTE from the consideration of metric spaces of weighted PINs and random walk walker distributions. We prove a theorem relating to convergence in these two metric spaces, before explaining how the metric from the space of walker distributions can be used to construct NTE. We subsequently validate the NTE measure on some simple synthetic networks and on a small biological data weighted PIN.
We next consider a separate entropy measure of our network traffic model, namely the entropy rate of a random walk on a weighted PIN, which we call signalling entropy. We briefly explain the history of this measure in a network theoretic context, noting that it has been demonstrated to be correlated with oncogenic status. We then explore two theoretical hypotheses regarding why signalling entropy may have this biological relevance. Firstly we demonstrate that under our model of network traffic, signalling entropy is most sensitive to the data values of highly connected vertices, in line with mutations in hub proteins driving high signalling entropy in cancer. Secondly, we show that signalling entropy, on average, is higher if computed over data samples which correspond to heterogeneous mixtures of different cell types rather than homogeneous cell types, in line with cancerous tissue being more heterogeneous than healthy tissue.
Finally we introduce the InSpiRe algorithm (based on a local vertex specific entropy) designed to compare two weighted PINs describing different biological phenotypes, in a manner which extracts a subset of proteins and interactions which is significantly rewiring across the phenotypes.
We close with a discussion motivating the development and application of these network theoretic tools performed in subsequent chapters.

2.1 A Random Walk Model of Network Traffic

The choice of a general dynamic model for a weighted network requires consideration of the literature. One must be careful to ensure the model makes minimal assumptions yet has sufficient descriptive power to portray complex systems. Much work has focused on interaction models known as *flow networks* (see for example [242]), in which transport from source vertices to sink vertices is subject to edge weight dependant constraints. These models are useful in optimisation problems where one wants to find paths through a network that maximise or minimise some function associated with path traversal, and thus tend to be used in systems where traffic can be manipulated, such as supply management [243].

Flow networks are less useful in the interrogation of network dynamics where constraints on traffic are unknown, and sink and source vertices and not readily defined. Moreover, when network dynamics are stochastic and bursty rather than continuous flows, such as in social communication systems [244] and gene regulatory networks [245], adaptations of flow networks are required. Such adaptations include discrete flow networks [246] and stochastic flow networks [247], in which the interactions of a vertex with its neighbours is given by a probability distribution proportional to the edge weight distribution. The elegance of these discrete models is that they may approximate continuous models (such as flow networks) in the large time limit. Such models are generalised, for example by the inclusion of holding rates in queuing theory [248], which with detailed information for parameter estimation can be used to describe and simulate a large variety of real world systems.

Given this literature we take our dynamic model for weighted networks as a balance between the descriptive power of the stochastic networks of queuing theory and the simplicity of the stochastic flow networks, namely a weighted random walk. The notion of random walks on graphs has a rich history [43, 249, 13], hence the novelty of our work does not derive from the nature of the model, but rather from its biological application and theoretical investigations. In what follows we introduce our model of biological network traffic.

Let $\mathcal{G} = (V, E)$ be an undirected graph, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices and $E = \{(i, j) | i, j \in V\}$ a set of edges; we denote the adjacency matrix of $\mathcal{G}$ by $A = (a_{ij})_{ij \in V}$. In our analysis $\mathcal{G}$ represents the undirected topology of the PIN, whence $a_{ij} = a_{ji}$.

For each biological sample considered we assign to each vertex $i \in V$ a variable $x_i \in \mathbb{R}^{>0}$, and denote the vector containing all such variables by $x = (x_i)_{i=1}^n \in \Omega \subset (\mathbb{R}^{>0})^n$, where Ω is some bounded domain. In our analysis x will represent a vector of gene/protein expression values, for a single sample.

We consider a random walk, on the graph $\mathcal{G}$; with transition probability matrix $P = (p_{ij})_{ij \in V}$ defined at either a phenotype level or a single sample level. In defining P we aim to model the interaction probabilities between connected proteins in the PIN.
At the single sample level, for each sample x we define

$$p_{ij} := p_{ij}(x) = \frac{a_{ij}x_j}{\sum_{k \in V} a_{ik}x_k}.$$

This choice of P is motivated by appealing to a simplified version of the mass action principle, namely that the rate of a reaction is proportional to the product of the active masses of the reagents involved. In this model we assume that normalised gene/protein expression is a rough proxy for protein concentration [23]. We stress that at the single sample level, p_{ij} is independent of x_i and that the only contribution of vertex i to the value of p_{ij} derives from the neighbourhood of i in the normalisation of p_{ij}.
At the phenotype level, given a set of samples corresponding to a phenotype $X := \{x^r\}_{r=1}^l$ we define the stochastic matrix $p_{ij} := p_{ij}^{pheno}$ either by

$$p_{ij}^{pheno} = \frac{a_{ij}(1 + cor(X_i, X_j))}{\sum_{k \in V} a_{ik}(1 + cor(X_i, X_k))},$$

or

$$p_{ij}^{pheno} = \frac{a_{ij}|cor(X_i, X_j)|}{\sum_{k \in V} a_{ik}|cor(X_i, X_k)|}.$$

Where $X_i := \{x_i^r\}_{r=1}^l$ and

$$cor(X_i, X_j) = \frac{\sum_{r=1}^l (x_i^r - \frac{1}{l}\sum_{r=1}^l x_i^r)(x_j^r - \frac{1}{l}\sum_{r=1}^l x_j^r)}{\sqrt{\sum_{r=1}^l (x_i^r - \frac{1}{l}\sum_{r=1}^l x_i^r)^2}\sqrt{\sum_{r=1}^l (x_j^r - \frac{1}{l}\sum_{r=1}^l x_j^r)^2}},$$

is the Pearson correlation between X_i and X_j. This choice of model is motivated by the assumption that signal transduction flows preferably along paths of proteins with correlated gene expression profiles.
The first version of this model, which favours positive correlations

$$p_{ij}^{pheno} = \frac{a_{ij}(1 + cor(X_i, X_j))}{\sum_{k \in V} a_{ik}(1 + cor(X_i, X_k))}$$

has been employed previously [23, 13]. It was noted, however, that this was only an approximation [23], and there is mounting evidence that negative correlations in gene expression play an important role in signal transduction [250], consequentially the model

$$p_{ij}^{pheno} = \frac{a_{ij}|cor(X_i, X_j)|}{\sum_{k \in V} a_{ik}|cor(X_i, X_k)|},$$

which assigns both strong positive and strong negative correlations a high weight may be considered more realistic. We note that unlike the single sample stochastic matrix, the

$(i,j)^{th}$ entry of the phenotype level matrix p_{ij}^{pheno}, is dependent on both X_i and X_j.
In the case of NTE the theoretical results derived in this chapter apply equally to any choice of stochastic matrix. For signalling entropy, we only consider the single sample stochastic matrix as this is the most analytically tractable for our purposes. For the phenotype level InSpiRe algorithm, we will explore the implications of using each of the two phenotype level stochastic matrices in Chapter 5 in the context of FSHD gene expression data.

2.2 Network Transfer Entropy and Metric Space

2.2.1 Introduction

In this section we will first investigate our probabilistic model of network traffic from the perspective of metric spaces. We will then utilise this global perspective to construct our local NTE measure.
We will consider an arbitrary edge weighted network, represented as a stochastic matrix as described above, equipped with a further vector of vertex weights or *initial signal distribution* (ISD). By examining the evolution of the ISD on the network according to a random walk determined by the edge weights, we will derive a closed form expression for the probability distribution of vertex weights after an arbitrary number of walker evolutions.
We will subsequently consider a metric space of signal dynamics, corresponding to the space of such probability distributions derived from weighted networks with the same ISD and topology, equipped with the square root *Jensen-Shannon Divergence* (JSD) metric. We prove a convergence principle, demonstrating that a form of convergence of weighted networks on N vertices (represented by stochastic matrices) in the metric space $L^p(\mathbb{M}^{N\times N})$ implies convergence in the constructed metric space of signal dynamics. This result shows that in our general framework, deformation of network structure influences network dynamics in an intuitive way. This result has real world implications in, for example, network drug design, where one wishes to modify the chemical affinities of interacting proteins in a pathological signalling network (*i.e.* modify the edge weights) to restore a healthy signalling regime (*i.e.* modify the dynamics).
Following this global theoretical investigation we consider an application of our metric space framework to derive a local measure of information transfer between two vertices in a weighted network (analogous to the transfer entropy measure for time series introduced by Schreiber [251]) which we name NTE. After constructing this measure we demonstrate its validity with some simple examples before applying our it to the quantification of in-

formation transfer between proteins in a phosphorylation network. Finally, we motivate potential further work in this area by explaining how theoretical problems in network evolution and network perturbation can be approached within our framework.

2.2.2 Information transfer on weighted networks and metric space

We here consider the above described Markovian model for signal dynamic evolution on a weighted network, in which each vertex is assigned a value quantifying a signal that the given vertex is capable of forwarding to one its neighbours. The vector containing these values for every vertex is the ISD of the network. In real world applications this distribution can be qualitatively diverse. For example, in biological networks, where vertices represent genes, the ISD may quantify the differential expression of genes in pathological versus healthy samples, or correlation between gene expression and clinical outcome. There are no restrictions on the ISD other than it being a vector in $\mathbb{R}^N$, where N is the number of vertices.

We evolve this signal over a network $\mathcal{G}$ with N vertices according to a stochastic matrix $P = (p_{ij})_{ij=1}^{N} \in \mathbb{M}^{N\times N}$, which may be derived from experimental and interaction data as described in Section 2.1. The ISD is propagated over the network in multiple discrete time steps. At each time step the signal at each vertex i is independently forwarded to vertex $j \in \mathcal{N}_i$ with probability p_{ij} (we emphasise that self-edges can be added to the network and the weights on such edges would determine the probability that a vertex maintains its signal over a single time step). Thus the number of time steps directly corresponds to the maximal path length the ISD has traversed. Given an ISD, $\vec{X}_0$, and path length, n, for every vertex i in the network we can compute the probability distribution of the signal at vertex i given that the ISD has been forwarded along paths of length n (*i.e.*, over n time steps), denoted $P[X_n^i|\vec{X}_0]$.

In this section we first derive a closed form expression for $P[X_n^i|\vec{X}_0]$. We then show how a metric space over such distributions for a given ISD and network topology can be constructed. Finally, we prove a convergence criterion relating this space to the metric space $L^p(\mathbb{M}^{N\times N})$.

2.2.2.1 A closed form expression of $P[X_n^i|\vec{X}_0]$

To derive a closed form expression for $P[X_n^i|\vec{X}_0]$, we first note that

$$P[X_n^i = y|\vec{X}_0] = \sum_{\vec{x}} P[\vec{X}_n = \vec{x}|\vec{X}_0]\delta_{x_i}^y, \tag{1}$$

Where δ_i^j denotes the Kronecker delta of i and j.
In addition by the Markovian nature of our dynamic model

$$P[\vec{X}_n = \vec{x}|\vec{X}_0] = \sum_{\vec{X}_1,\ldots,\vec{X}_{n-1}} P[\vec{X}_n = \vec{x}|\vec{X}_{n-1}]\ldots P[\vec{X}_1|\vec{X}_0], \tag{2}$$

Reducing our problem to the calculation of the transition probabilities $P[\vec{X}_{k+1}|\vec{X}_k]$, between states and the states themselves which must be summed over. These are not, however, immediate.
Let us first consider the following: given we know the full signal distribution at time-point $k \in \mathbb{N}$ *i.e.* $\vec{X}_k = \vec{x}_k$, then all possible states of signal distribution at time-point $k+1$ have the form

$$\vec{X}_{k+1} = A^T\vec{x}_k. \tag{3}$$

Here $A = (A_{ij})_{i,j=1}^N$, $A_{ij} \in \{0,1\}$ is a binary matrix with a single non-zero entry in every row; the column index j of the non-zero entry in row i corresponds to the unique vertex j that i has sent its signal to during the time-step $k \to k+1$. We note that in addition $A_{ij} = 0$ if $j \notin \mathcal{N}_i$ and that A is independent of $\vec{x}_k$.
Thus every realisation of a single signal transfer event in a given weighted network can be represented as a matrix operation A, independently of ISD. We denote the set of such matrices by $\mathcal{A}$, and emphasise that it depends only upon the topology (*i.e.*, zero pattern) of the weighted network.
It is clear that for $N < \infty$ the set $\mathcal{A}$ must be countable, and its cardinality must be $|\mathcal{A}| = \prod_{i=1}^N k_i$, where $k_i = |\mathcal{N}_i|$. Moreover, it is clear that we can construct every element in $\mathcal{A}$ given $\cup_{i=1}^N \mathcal{N}_i$.
Following this, it is clear that given any ISD $\vec{X}_0$ the signal distribution at time point $k > 0$ must have the (possibly non-unique) form

$$\vec{X}_k = A_k^T\ldots A_1^T\vec{X}_0, \tag{4}$$

where $A_i \in \mathcal{A}$ for $i = 1,\ldots,k$. Whence Eq. (2) can be expressed as

$$P[\vec{X}_n = \vec{x}|\vec{X}_0] = \sum_{A_1,\ldots,A_{n-1}\in\mathcal{A}} P[\vec{X}_n = \vec{x}|\vec{X}_{n-1} = A_{n-1}^T\ldots A_1^T\vec{X}_0]\ldots P[\vec{X}_1 = A_1^T\vec{X}_0|\vec{X}_0]. \tag{5}$$

Thus to compute $P[X_n^i|\vec{X}_0]$, it suffices to compute

$$P[\vec{X}_{k+1} = A_{k+1}^T\ldots A_1^T\vec{X}_0|\vec{X}_k = A_k^T\ldots A_1^T\vec{X}_0],$$

which is simply the probability of the signal dynamic A_{k+1} being selected from $\mathcal{A}$. By model construction this can be expressed as

$$\prod_{i=1}^{N}\sum_{j=1}^{N} p_{ij}A_{ij}.$$

Whence combining this with (1) and (5) we derive the closed form expression

$$P[X_n^i = y|\vec{X_0}] = \sum_{A_1,\ldots,A_n\in\mathcal{A}} \delta^y_{(A_1^T\ldots A_n^T\vec{X}_0)_i} \prod_{r=1}^{n}\prod_{k=1}^{N}\sum_{j=1}^{N} p_{kj}(A_r)_{kj}. \tag{6}$$

2.2.2.2 Metric Space We demonstrated above how, for a specific network $\mathcal{G}$, and ISD $\vec{X}_0$ we can calculate a set of matrices describing possible signal dynamics over a single time-step of our model, as well as a probability distribution describing the signal on the entire network after n time steps. We will denote these constructs for the weighted network $\mathcal{G}$ by $\mathcal{A}_\mathcal{G}$ and $P_\mathcal{G}[\vec{X}_n|\vec{X}_0]$, respectively and stress that the former only depends on the topology of $\mathcal{G}$ and not the edge weights, we will denote the topology of $\mathcal{G}$ by $t(\mathcal{G})$ (as noted before, topology in this context refers to the zero pattern of the network and is independent of the edge weights). The probability distribution $P_\mathcal{G}[\vec{X}_n|\vec{X}_0]$ is a measure over the finite set

$$\{A_1^T\ldots A_n^T\vec{X}_0 : (A_i)_{i=1}^n \subset \mathcal{A}_\mathcal{G}\},$$

which we will denote $\Omega^n_{\vec{X}_0}(t(\mathcal{G}))$. If we denote the space of probability measures over $\Omega^n_{\vec{X}_0}(t(\mathcal{G}))$ by $M_1^+(\Omega^n_{\vec{X}_0}(t(\mathcal{G})))$, then it is clear that for any two weighted networks $\mathcal{G}_1$ and $\mathcal{G}_2$ with the same topology, T, the probability distributions $P_{\mathcal{G}_1}[\vec{X}_n|\vec{X}_0]$ and $P_{\mathcal{G}_2}[\vec{X}_n|\vec{X}_0]$ are elements of $M_1^+(\Omega^n_{\vec{X}_0}(T))$.

It has been shown that for any measurable space Ω, the square root of the JSD induces a metric on the space $M_1^+(\Omega)$ [252]. The JSD is defined by

$$D_{JS}(p,q) = \frac{1}{2}\left(\sum_{x\in\mathcal{X}}\left(p(x)\log\frac{p(x)}{m(x)} + q(x)\log\frac{q(x)}{m(x)}\right)\right) \tag{7}$$

where $p, q : \mathcal{X} \to [0,1]$ are probability distributions (with no restrictions placed on their kernels) and $m = (p+q)/2$. Thus the quantity

$$\sqrt{D_{JS}(P_{\mathcal{G}_1}[\vec{X}_n|\vec{X}_0], P_{\mathcal{G}_2}[\vec{X}_n|\vec{X}_0])}$$

computes a metric distance between the probability distributions describing the dynamics on $\mathcal{G}_1$ and $\mathcal{G}_2$.

Thus our network traffic framework results in a mapping from the space of weighted networks to a family of metric spaces in which elements of the metric space represent possible signal dynamics.

2.2.2.3 Convergence Principle We have above constructed a family of mappings from the space of weighted networks to a family of metric spaces, in which elements of the metric spaces correspond to signal dynamics on the networks. The mappings and structure of the metric spaces are parametrised by the path length parameter n, the ISD $\vec{X}_0$ and the topology (*i.e.,* the zero pattern, but not the edge weights) of the network. This formalism allows a more theoretical treatment of dynamics on networks from the perspective of metric spaces, and permits a coupling between weighted network structure and dynamics. In certain fields, understanding the reaction of network dynamics to perturbations of edge weights is of great importance. This is particularly true of network drug design [253], in which one is interested in sequentially deforming the quantitative strengths of interactions in a pathological signalling network (via drugs) into those of a healthy network, with the aim of establishing a healthy gene expression dynamic and mitigating the pathology. If this notion of treatment is logical within our framework, then one would postulate that convergence in weight distribution of a sequence of networks to a limit distribution (in a matrix metric space) would imply convergence of the corresponding sequence of dynamics to the dynamics of the limit network (in the network dynamic metric space). We thus consider the following theorem

Theorem 1 (Convergence Principle) *Let $(W^n)_{n\in\mathbb{N}}$ be a sequence of weighted networks on N vertices of fixed topology, T, and let $(P^m)_{m\in\mathbb{N}} \subset [0,1]^{N\times N}$ be the corresponding row normalised stochastic matrices for the sequence. Let $P \subset [0,1]^{N\times N}$ be a stochastic matrix of topology T. If $P^m \to P$ in $L^p(\mathbb{M}^{N\times N})$, $p \geq 1$, as $m \to \infty$, then for fixed ISD $\vec{X}_0$ and path length parameter n, the signal distributions*

$$P_{P^m}[\vec{X}_n|\vec{X}_0] \to P_P[\vec{X}_n|\vec{X}_0]$$

as $m \to \infty$ in the metric space

$$\left(M_1^+(\Omega^n_{\vec{X}_0}(T)), \sqrt{D_{JS}(\cdot,\cdot)}\right).$$

Proof: Firstly we define the L^p norm of a matrix $A \in \mathbb{M}^{N\times N}$

$$||A||_p := \begin{cases} \left(\sum_{i,j=1}^N |a_{ij}|^p\right)^{1/p} & \text{if } p < \infty \\ \max_{i,j} |A_{ij}| & \text{if } p = \infty \end{cases}. \tag{8}$$

A well-known and easy to derive bound on L^p spaces, which holds for any $A \in \mathbb{M}^{N\times N}$ is $||A||_\infty \leq ||A||_p$.
Fix $\epsilon > 0$. As $P^m \to P$ in $L^p(\mathbb{M}^{N\times N})$, there exists $M \in \mathbb{N}$ such that for all $m \geq M$,

$$||P^m - P||_p \ < \epsilon.$$

Let us define the matrix $\Delta^P \in (-1,1)^{N\times N}$ via

$$\Delta^P := P^M - P,$$

it is clear that $||\Delta^P||_\infty < \epsilon$.
We now consider for a fixed ISD $\vec{X}_0$ and path length parameter n the distributions $P_P[\vec{X}_n = \vec{x}|\vec{X}_0]$ and $P_{P^M}[\vec{X}_n = \vec{x}|\vec{X}_0]$, which we will hereafter refer to as $\mathbb{P}_P(\vec{x})$ and $\mathbb{P}_{P^M}(\vec{x})$. It was shown above that

$$\mathbb{P}_P(\vec{x}) = \sum_{A_1,\ldots,A_n\in\mathcal{A}} \delta^{\vec{x}}_{(A_1^T\ldots A_n^T\vec{X}_0)} \prod_{r=1}^{n}\prod_{i=1}^{N}\sum_{j=1}^{N} P_{ij}(A_r)_{ij}. \tag{9}$$

The set of possible signal dynamics $\mathcal{A}$ for a given weighted network was also explicitly constructed above and was shown to depend only on the network topology and not on the edge weights of the network. Consequentially as the sequence $(P^n)_{n\in\mathbb{N}}$ and the network P have the same topology T, the set of possible signal dynamics $\mathcal{A}$ is the same for every element of the sequence and the network P.
Consider expanding the product

$$\begin{aligned}
\prod_{r=1}^{n}\prod_{i=1}^{N}\sum_{j=1}^{N} P_{ij}(A_r)_{ij} &= \prod_{r=1}^{n}\prod_{i=1}^{N}\sum_{j=1}^{N}(P^M_{ij} - \Delta^P_{ij})(A_r)_{ij} \\
&= \Big[\big(\sum_{j=1}^{N}(P^M_{1j}(A_1)_{1j} - \sum_{j=1}^{N}\Delta^P_{1j})(A_1)_{1j}\big)\ldots \\
&\quad \big(\sum_{j=1}^{N}(P^M_{Nj}(A_1)_{Nj} - \sum_{j=1}^{N}\Delta^P_{Nj})(A_1)_{Nj}\big)\Big] \\
\ldots &\quad \Big[\big(\sum_{j=1}^{N}(P^M_{1j}(A_n)_{1j} - \sum_{j=1}^{N}\Delta^P_{1j})(A_n)_{1j}\big)\ldots \\
&\quad \big(\sum_{j=1}^{N}(P^M_{Nj}(A_n)_{Nj} - \sum_{j=1}^{N}\Delta^P_{Nj})(A_n)_{Nj}\big)\Big].
\end{aligned} \tag{10}$$

Grouping together terms we can express the product as

$$\begin{aligned}\prod_{r=1}^{n}\prod_{i=1}^{N}\sum_{j=1}^{N}P_{ij}(A_r)_{ij} &= \prod_{r=1}^{n}\prod_{i=1}^{N}\sum_{j=1}^{N}P^M_{ij}(A_r)_{ij} \\ &- \left[\sum_{i=1}^{N}\sum_{r=1}^{n}\Big(\sum_{j=1}^{N}\Delta^P_{ij}(A_r)_{ij}\Big)\prod_{l\neq r}^{n}\prod_{s\neq i}^{N}\Big(\sum_{j=1}^{N}P^M_{sj}(A_l)_{sj}\Big)\right] \\ &+o(\epsilon). \end{aligned} \tag{11}$$

We will denote the second term in the above expression by

$$\mathcal{H}_{(A_i)_{i=1}^n}(\Delta^P) := \left[\sum_{i=1}^{N}\sum_{r=1}^{n}\Big(\sum_{j=1}^{N}\Delta^P_{ij}(A_r)_{ij}\Big)\prod_{l\neq r}^{n}\prod_{s\neq i}^{N}\Big(\sum_{j=1}^{N}P^M_{sj}(A_l)_{sj}\Big)\right] \tag{12}$$

Substitution of (11) into (9) yields

$$\mathbb{P}_P(x) = \sum_{A_1,\ldots,A_n\in\mathcal{A}}\delta^{\vec{x}}_{(A_1^T\ldots A_n^T\vec{X}_0)}\left[\prod_{r=1}^{n}\prod_{i=1}^{N}\sum_{j=1}^{N}P^M_{ij}(A_r)_{ij} - \mathcal{H}_{(A_i)_{i=1}^n}(\Delta^P) + o(\epsilon)\right]. \tag{13}$$

Clearly from (9) the first term can be expressed

$$\sum_{A_1,\ldots,A_n\in\mathcal{A}}\delta^{\vec{x}}_{(A_1^T\ldots A_n^T\vec{X}_0)}\prod_{r=1}^{n}\prod_{i=1}^{N}\sum_{j=1}^{N}P^M_{ij}(A_r)_{ij} = \mathbb{P}_{P^M}(x). \tag{14}$$

For the second term, notice that

$$\begin{aligned}\left|\sum_{A_1,\ldots,A_n\in\mathcal{A}}\delta^{\vec{x}}_{(A_1^T\ldots A_n^T\vec{X}_0)}\mathcal{H}_{(A_i)_{i=1}^n}(\Delta^P)\right| &\leq \sum_{A_1,\ldots,A_n\in\mathcal{A}}\delta^{\vec{x}}_{(A_1^T\ldots A_n^T\vec{X}_0)}\sum_{i=1}^{N}\sum_{r=1}^{n} \\ &\Big(\sum_{j=1}^{N}||\Delta^P||_\infty(A_r)_{ij}\Big)\prod_{l\neq r}^{n}\prod_{s\neq i}^{N}\Big(\sum_{j=1}^{N}P^M_{sj}(A_l)_{sj}\Big) \\ &< \epsilon\sum_{A_1,\ldots,A_n\in\mathcal{A}}\delta^{x}_{(A_1^T\ldots A_n^T\vec{X}_0)}\sum_{i=1}^{N}\sum_{r=1}^{n} \\ &\prod_{l\neq r}^{n}\prod_{s\neq i}^{N}\Big(\sum_{j=1}^{N}P^M_{sj}(A_l)_{sj}\Big) \\ &\leq \left(\prod_{i=1}^{N}k_i\right)^n nN\epsilon \end{aligned} \tag{15}$$

where the second inequality follows from $||\Delta^P||_\infty < \epsilon$ and $\sum_{j=1}^{N}(A_r)_{ij} = 1$ by construction of the set $\mathcal{A}$. The final inequality follows from the facts that $\sum_{j=1}^{N}P^M_{sj}(A_l)_{sj} \leq 1$ and $|\mathcal{A}| = \prod_{i=1}^{N}k_i$. Given these bounds and Eq. (13) it follows that

$$\mathbb{P}_P(x) < \mathbb{P}_{P^M}(x) + \left(\prod_{i=1}^{N}k_i\right)^n nN\epsilon + o(\epsilon). \tag{16}$$

An identical argument can be used exchanging P^M and P, in which case the sign of $\mathcal{H}_{(A_i)_{i=1}^n}(\Delta^P)$ in (13) changes to positive, however the bound established in (15) bounds the modulus of $\mathcal{H}_{(A_i)_{i=1}^n}(\Delta^P)$ and thus will always be greatest than the largest negative or largest positive value of $\mathcal{H}_{(A_i)_{i=1}^n}(\Delta^P)$. Thus we obtain the symmetric bound

$$\mathbb{P}_{P^M}(x) < \mathbb{P}_P(x) + \left(\prod_{i=1}^{N} k_i\right)^n nN\epsilon + o(\epsilon). \tag{17}$$

Let us define

$$m(x) := \frac{1}{2}\left(\mathbb{P}_P(x) + \mathbb{P}_{P^M}(x)\right),$$

it follows from (16) and (17) that

$$\frac{\mathbb{P}_P(x)}{m(x)} < \frac{2\mathbb{P}_P(x)}{2\mathbb{P}_P(x) - \left(\prod_{i=1}^{N} k_i\right)^n nN\epsilon + o(\epsilon)} \tag{18}$$

and

$$\frac{\mathbb{P}_{P^M}(x)}{m(x)} < \frac{2\mathbb{P}_{P^M}(x)}{2\mathbb{P}_{P^M}(x) - \left(\prod_{i=1}^{N} k_i\right)^n nN\epsilon + o(\epsilon)}. \tag{19}$$

Thus it follows that

$$\begin{aligned} D_{JS}(\mathbb{P}_P, \mathbb{P}_{P^M}) &= \frac{1}{2}\left(\sum_x \mathbb{P}_P(x)\log\frac{\mathbb{P}_P(x)}{m(x)} + \mathbb{P}_{P^M}(x)\log\frac{\mathbb{P}_{P^M}(x)}{m(x)}\right) \\ &< \frac{1}{2}\left(\sum_x \log\left(\frac{2\mathbb{P}_P(x)}{2\mathbb{P}_P(x) - \left(\prod_{i=1}^{N} k_i\right)^n nN\epsilon + o(\epsilon)}\right)\right. \\ &+ \left.\log\left(\frac{2\mathbb{P}_{P^M}(x)}{2\mathbb{P}_{P^M}(x) - \left(\prod_{i=1}^{N} k_i\right)^n nN\epsilon + o(\epsilon)}\right)\right). \end{aligned} \tag{20}$$

By algebra of limits, it is clear that as $\epsilon \to 0$

$$\frac{2\mathbb{P}_P(x)}{2\mathbb{P}_P(x) - \left(\prod_{i=1}^{N} k_i\right)^n nN\epsilon + o(\epsilon)} \to 1$$

and

$$\frac{2\mathbb{P}_{P^M}(x)}{2\mathbb{P}_{P^M}(x) - \left(\prod_{i=1}^{N} k_i\right)^n nN\epsilon + o(\epsilon)} \to 1.$$

Whence it follows that

$$D_{JS}(\mathbb{P}_P, \mathbb{P}_{P^m}) \to 0$$

as $m \to \infty$ and the theorem is true.

□

We note that the theorem also holds if the topologies of the sequence of weighted networks and limit network are different from one another, the proof of this statement follows precisely as above, the only difference being the set $\mathcal{A}$ utilised is that induced by the complete graph topology.

2.2.3 Network Transfer Entropy

In the previous section we demonstrated how our random walk model of network traffic finds a natural interpretation in the concepts of metric space. We further demonstrated that in this framework global network dynamics are influenced by global network structure in a natural way. In this section we consider our random walk model in a more local setting to construct an information theoretic measure, NTE, quantifying the directed amount of information transferred between any two vertices in a weighted network. Following construction, we demonstrate our measure on simple synthetic networks and a biological signalling network.

Transfer entropy was a measure introduced by Schreiber, to quantify the directed amount of information transferred between two mutually dependent time series [251] and is an important starting point when considering the construction of our own measure of information transfer. In what follows we employ an approach analogous to Schreiber's to construct NTE.

The definition of Schreiber's transfer entropy required a model in which it was possible to express whether two time series influenced each other. For transfer entropy to be widely applicable, this model needed to be sufficiently general to portray a wide array of diverse systems. It was thus intuitive to describe time series as realisations of (approximately) Markov processes of order k. For such a process I the conditional probability of finding the process in state i_{n+1} at time $n+1$ satisfies

$$p(i_{n+1}|i_n,\ldots,i_{n-k+1}) = p(i_{n+1}|i_n,\ldots,i_{n-k}). \tag{21}$$

These generalised Markov process are not all encompassing in their descriptive power; for example, they are in general not-applicable to studying subsystems of Markov processes [254]. However for a broad range of datasets including heart and breathing rate data [251], magnetoencephalography data [255] and financial time series [256], the approximate Markov process model can be justified, making Schreiber's transfer entropy widely applicable.

It is clear that our Markovian model of signal transfer can be analogised to that of Schrieber and hence is suitable for construction of a network transfer entropy. For our

purposes we wish to identify the directed amount of information transferred between any two pairs of network vertices during a period of system evolution. In the derivation of Schreiber's transfer entropy the directed amount of information transferred from a process J to a process I is formulated as the incorrectness of the assumption that I is not conditional on J. This quantity can be expressed as the *Kullback-Leibler divergence* [257]

$$\sum p(i_{n-k+1}^{n+1}, j_l^n) \log \frac{p(i_{n+1}|i_{n-k+1}^n, j_l^n)}{p(i_{n+1}|i_{n-k+1}^n)}, \tag{22}$$

where $i_m^n = (i_n, \ldots, i_m)$ for $m < n$. Thus to quantify the amount of information vertex j in our network transfers to vertex i over paths of length n we must derive a distribution for X_n^i (the signal at vertex i after n signal transfer events) in which vertex j sends no information to vertex i. We must then compute the informational distance between this distribution and the above considered distribution $P[X_n^i|\vec{X}_0]$ in which j is able to communicate with i. Clearly if we set the j^{th} row in the stochastic matrix P to $\vec{e}_j$ (*i.e.*, make j an absorbing state; $\vec{e}_j$ denotes the j^{th} element of the standard basis of $\mathbb{R}^N$), then it is impossible for vertex j to communicate with any vertex $i \neq j$ under our model. Given this modified matrix we can compute the probability distribution of X_n^i given the ISD and that j cannot communicate with i. We denote this distribution $P[X_n^i|\vec{X}_0, j]$. Here we diverge from Schreiber, however, as the Kullback-Leibler divergence

$$\sum P[X_n^i|\vec{X}_0] \log \frac{P[X_n^i|\vec{X}_0]}{P[X_n^i|\vec{X}_0, j]}$$

is only well defined provided

$$\{x : P[X_n^i = x|\vec{X}_0, j] = 0\} \subset \{x : P[X_n^i = x|\vec{X}_0] = 0\},$$

which is an assumption that does not hold in general. Consider, for example, a directed graph on two vertices 1 and 2, with a single directed edge oriented from 1 to 2; if we assign the ISD as $\vec{X}_0 = \vec{e}_1$, then it is trivial that $P[X_1^2 = x|\vec{X}_0] = \delta_x^1$ and $P[X_1^2 = x|\vec{X}_0, 1] = \delta_x^0$. Thus to quantify the directed amount of information transferred from vertex j to i we must employ a different measure of statistical distance. There are several possible choices available, among the most promising are the previously considered JSD, which is a linear combination of Kullback-Leibler divergences, and the statistical distance introduced by Wootters [258]. Both measures are theoretically rich; Wootters' measure was designed as a distinguishably distance between pure quantum states after a finite number of observations, and applies equally well to distinguishing two probability distributions. The measure also has a geometric interpretation in the context of Hilbert space. The JSD of two distributions, quantifies the total Kullback-Leibler divergence from each distribution

to the average of the two, and thus is a measure of distributional similarity. As explored above, however, the JSD is also the square of a metric over the space of probability distributions on a measurable set [252] and it is worth mentioning also that the JSD and Wootters' measure have been shown to agree to second order in a quantum mechanical framework [259].

Given the relationship to our theoretical work in metric spaces, it is natural to use the JSD. Whence we define the *network transfer entropy* (NTE) from j to i over path length n and given an ISD $\vec{X}_0$ by

$$\tau^n_{\vec{X}_0}(j||i) := D_{JS}(P[X^i_n|\vec{X}_0], P[X^i_n|\vec{X}_0, j]) \tag{23}$$

Note that $\tau^n_{\vec{X}_0}(j||i) \in [0, \log 2]$ is inherently asymmetric, and thus quantifies information transfer through a network in a directed sense, permitting the inference of causality.

2.2.4 Simple examples

In order to demonstrate the use of NTE we first consider two simple synthetic networks as examples. To evaluate NTE in these examples, we estimated the probability distributions $P[X^i_n|\vec{X}_0]$ and $P[X^i_n|\vec{X}_0, j]$ for all j, using simulation, as described in the Materials and Methods, Chapter 2.

The first and most simple example we consider is a directed path of length 5 with equal edge weights (Fig 2.1). This induces the stochastic matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{24}$$

The structure of this network provides a completely predictable path for signal transfer and thus is ideal for proving the capability of our measure to detect information transfer. We consider two ISDs on this network, firstly $\vec{X}_0 = \vec{e}_1$, where the first vertex in the path is given an initial signal and all other vertices have no signal to transfer. If we number the vertices 1 to 5 from the start of the path to the end, then it is clear that for this ISD, over paths of length $n = 1$, vertex 1 sends information to vertex 2, and no other vertices communicate, for $n = 2$ vertex 1 sends information to vertex 3 and vertex 2 sends information to vertex 3 and no other vertices communicate, and similarly we can compute all pairwise information transfer events up to $n = 4$ beyond which all signal is absorbed at vertex 5 and cannot be transmitted through the network. This pattern is precisely

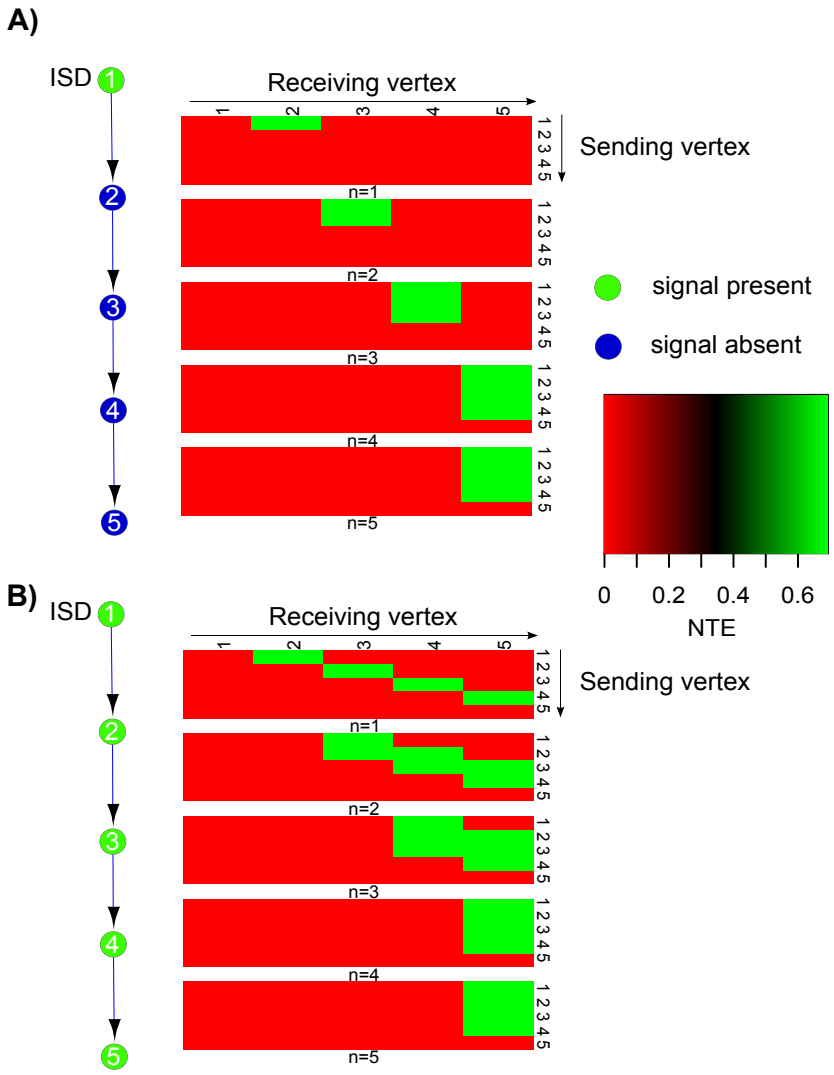

Figure 2.1: **NTE Example 1: Deterministic Path.** Matrices showing NTE between all vertex pairs in a deterministic path over path lengths $n = 1 - 5$ for ISDs (A) $(1, 0, 0, 0, 0)^T$ and (B) $(1, 1, 1, 1, 1)^T$.

what is seen if we calculate the NTE between all vertex pairs over different path lengths n (Fig 2.1A).

We next consider the ISD $\vec{X}_0 = (1,1,1,1,1)^T$, on the same network, in order to demonstrate the ability of the NTE measure to discern between situations where networks with identical edge weights have different starting signal states. One would expect that with this ISD, for $n = 1$, rather than just vertex 1 forwarding information to vertex 2, we have vertex j forwarding information to vertex $j+1$ for $j = 1,\ldots,4$, and similar extensions for longer path lengths. We find that the NTE measure can detect these differences due to ISD (Fig 2.1B).

The next network we consider is a slight extension to the deterministic path which constitutes a directed feedback from vertex 2 to vertex 4 weighted $w = x/(1-x)$ (Fig 2.2). This induces the stochastic matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & x & 0 & 0 & 1-x \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{25}$$

The network introduces some indeterminism in that if vertex 4 holds signal, it can either forward it to vertex 5 with probability $1-x$ or feedback the signal to vertex 2 with probability x. This essentially sets up a feedback loop which dampens the signal received at node 5 over a given path length, by a factor dependant on x. We calculated the NTE for all vertex pairs in this altered path for $x = 0.1, 0.5, 0.9$, corresponding to $w = 1/9, 1, 9$, and found that as x is increased the NTE to vertex 5 from all other vertices falls, as expected (Fig 2.2). Thus in the context of these very simple synthetic networks, the use of NTE as a tool for detecting information transfer is clear.

2.2.5 NTE on a biological network

We next consider NTE in the context of a real world biological network, specifically the human primary naive *CD4+* T cell intracellular signalling network analysed by Sachs *et al.* [260], consisting of 11 vertices (Fig 2.3). In this network vertices are proteins which can be phosphorylated and directed edges connect kinases (capable of phosphorylating proteins) with their targets. The kinases must be in an active state before they can phosphorylate a target; activity can be achieved by either phosphorylation by an upstream kinase or activation by a reagent. Sachs *et al.* generated data accompanying this network consisting of quantification (by flow cytometry) of the amount of phosphorylated protein

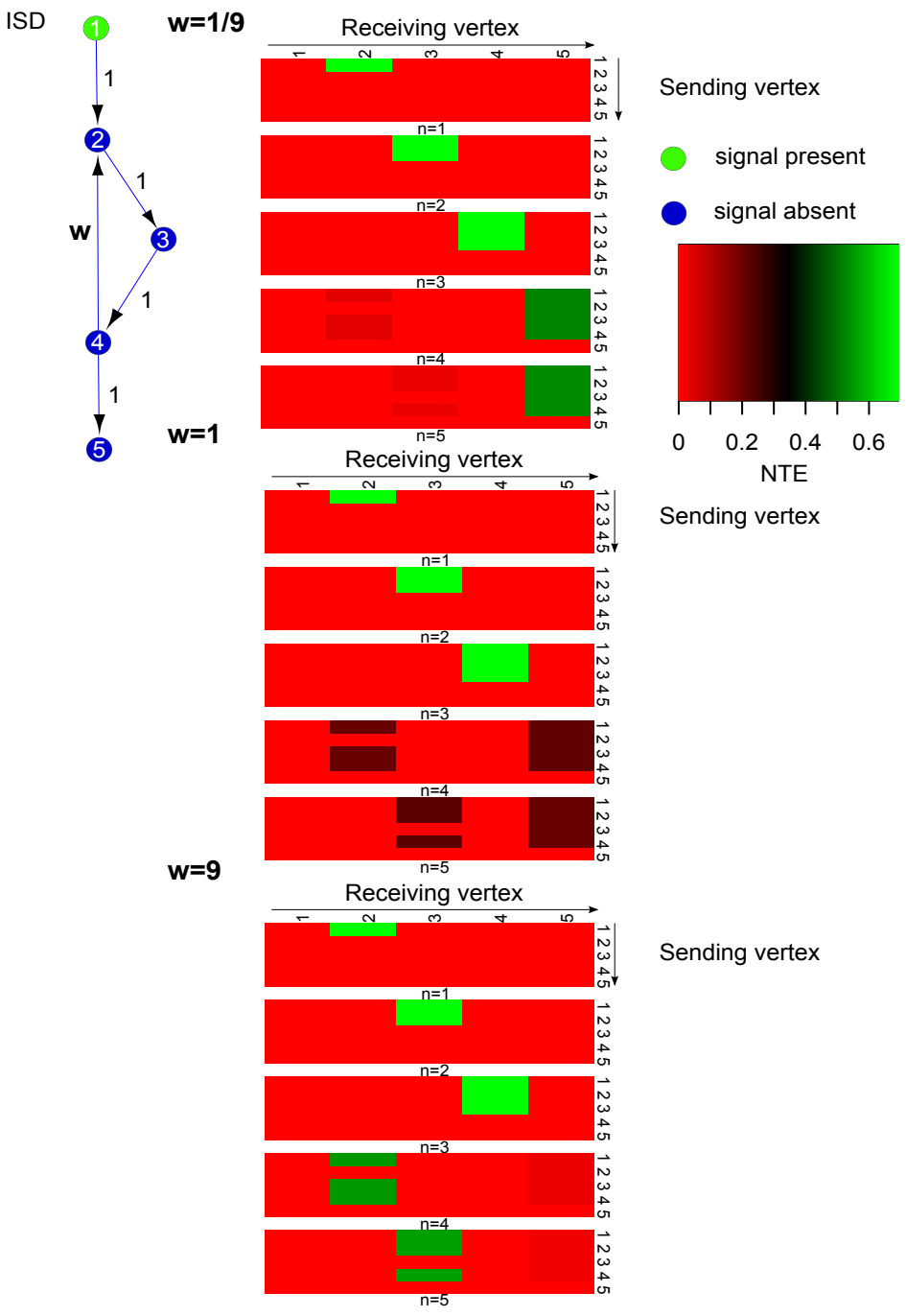

Figure 2.2: **NTE example 2: Directed Feedback.** Matrices showing NTE between all vertex pairs in the modified, weighted path with feedback from vertex 4 to vertex 2, over path lengths $n = 1 - 5$, for a range of feedback strengths. Note that as the weight on the edge $(4, 2)$ rises the NTE from vertices 1-4 to 5 falls.

for each vertex in the network, following ten independent perturbations. The network perturbations consisted of the administration of reagents which could either activate or inhibit the kinase activity of particular vertices.

To apply NTE we considered two of these perturbations, firstly treatment with anti-*CD3* and anti-*CD28* to activate the T-cells and induce flux through the network, secondly treatment with anti-*CD3*, anti-*CD28* and psitectorigenin, a reagent which specifically inactivates *PIP2*. The computation of NTEs on the perturbation of the biological network requires the assignment of network edge weights and an ISD for each perturbation, this requires careful consideration. The assigned weightings are presented in Fig 2.3A-B and are described in detail in the Materials and Methods, Chapter 2.

Following the assignment of ISDs and edge-weights to the two perturbations of the biological network, we computed matrices of NTEs between all vertex pairs for path lengths $n = 1 \ldots 5$ in each perturbation (Fig 2.3C-D).

To compare information transfer under the two perturbations we next considered the differential NTE evaluated from subtracting the *PIP2* inhibited (psitectorigenin treated) NTEs from the uninhibited NTEs (Fig 2.4). We found that in the psitectorigenin treated network, information transfer from *PIP2* to the rest of the network was reduced over all path lengths (Fig 2.4). Specifically, information transfer from *PIP2* to *PIP3* was greatly reduced and information transfer from *PIP2* to *Plcg* was reduced over paths of maximal length greater than one (implying *PIP3* received less information from *Plcg* via *PIP2* under psitectorigenin treatment). At longer path lengths we also see a reduced information transfer from *PIP2* to *Akt* and p38 in the psitectorigenin treated network. This indicates that specific inhibition of *PIP2* can lead to decreased *Akt* and p38 activation downstream of *PIP2* signalling.

Interestingly, we also notice that in the *PIP2* inhibited network, there is increased information transfer from *PKA* to *Akt* and from *PKC* to p38. This points at a compensatory mechanism, in which inhibition of *PIP2* leading to reduced *Akt* and p38 activation is compensated for by *PKC* dependant p38 activation and *PKA* dependant *Akt* activation. Thus our NTE measure is capable of providing novel insights into signalling mechanisms in biological networks.

2.2.6 Conclusions and Possible Further Work

In this subsection on NTE and metric space, we first constructed a general metric space framework for dynamics on weighted networks and proved a convergence principle relating global weighted network structure to dynamics. We next derived a general, local information theoretic measure, NTE, for quantifying the amount of information trans-

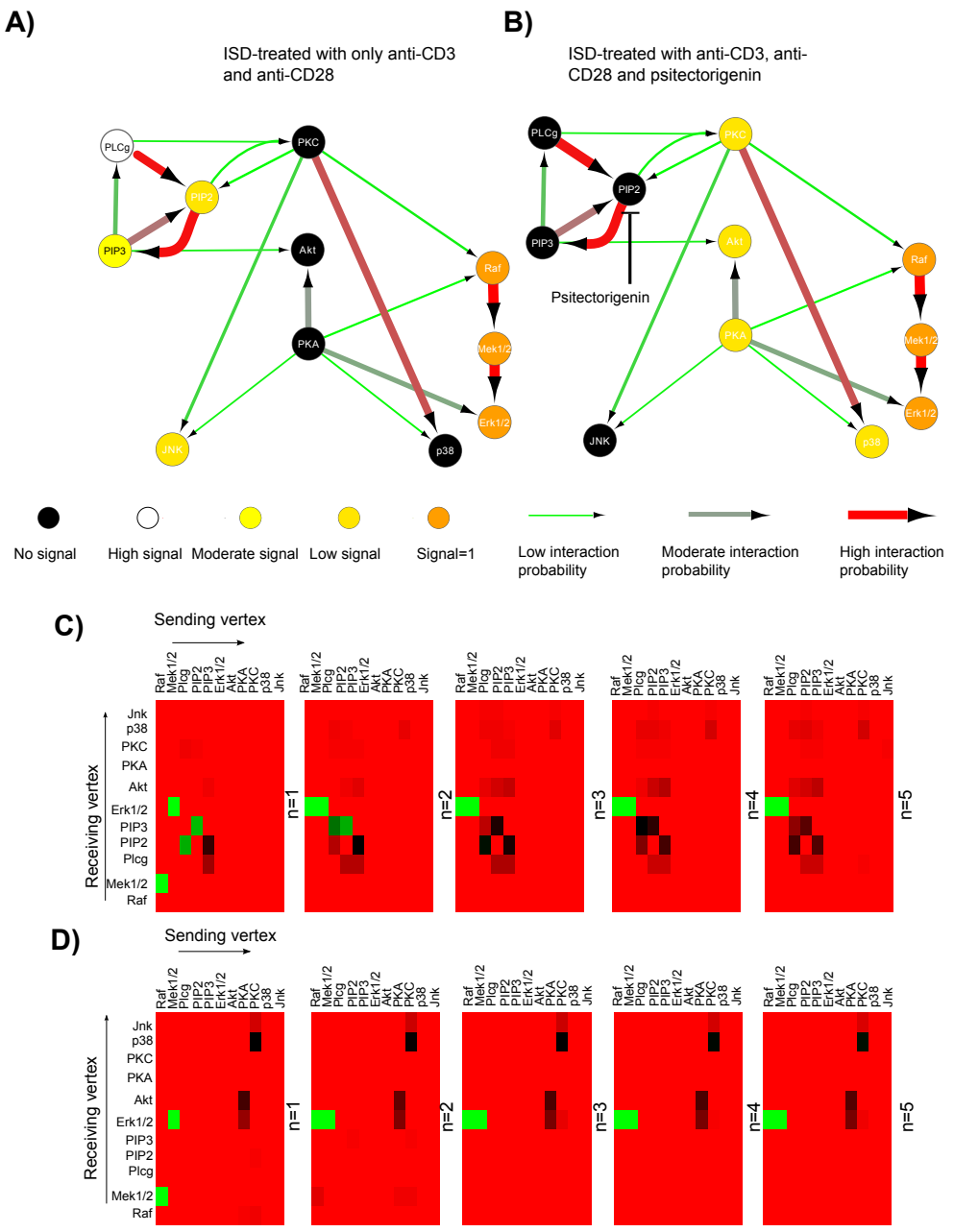

Figure 2.3: **ISDs, edge-weights and NTE computations on the biological network.** The top half of the figure shows the ISD and edge weights for (A) the anti-*CD3* and anti-*CD28* treated network and for (B) the anti-*CD3*, anti-*CD28* and psitectorigenin treated network. The bottom half displays matrices of NTEs computed between every vertex pair over path lengths $n = 1 - 5$, in (C) the anti-*CD3* and anti *CD28* treated network and (D) the anti-*CD3*, anti-*CD28* and psitectorigenin treated network.

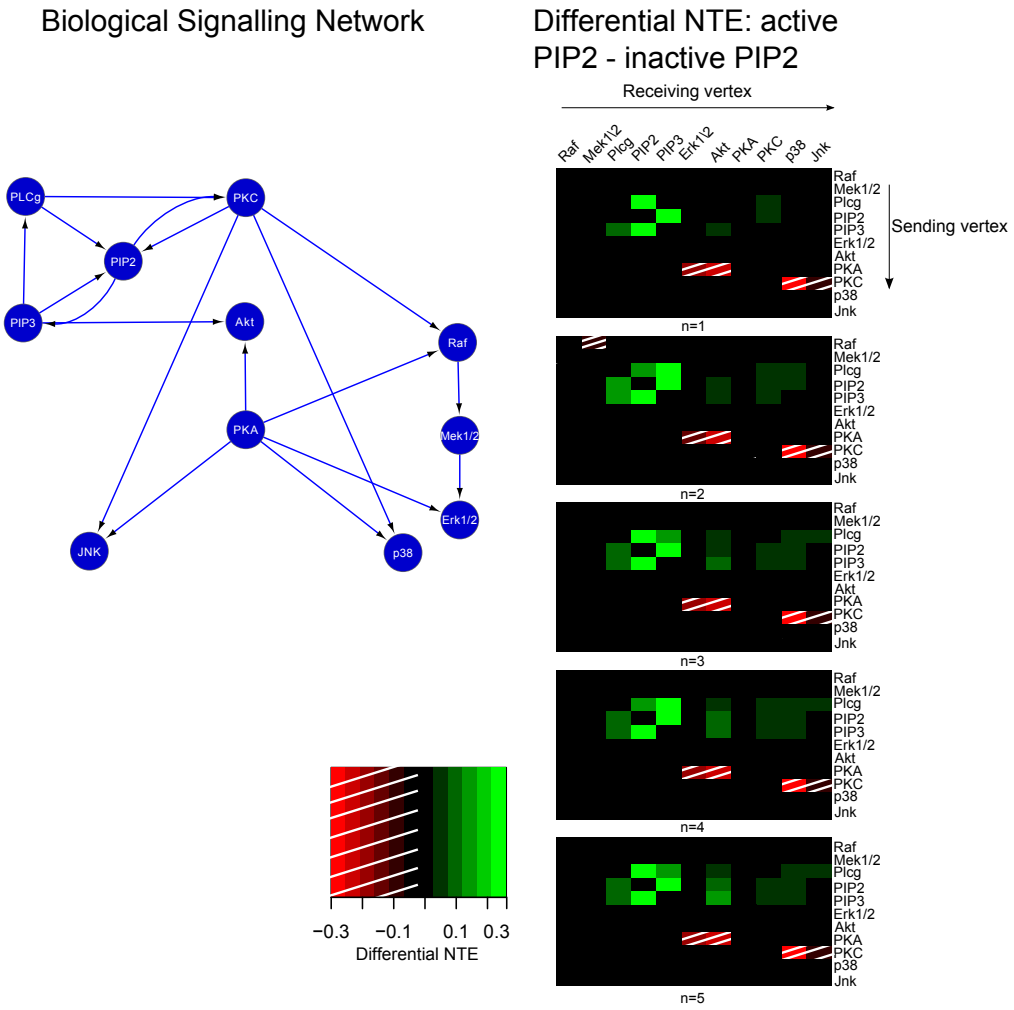

Figure 2.4: **NTE computed over two perturbations of a biological network.** Matrix displaying differential NTEs computed for every vertex pair over the displayed network, positive values (green) correspond to NTEs higher in the network perturbed with anti-*CD3* and anti-*CD28*, whilst negative values (red dashed) are higher in the network also perturbed by psitectorigenin, a *PIP2* inhibitor.

ferred between any two vertices of a weighted network over paths of varying length. We demonstrated our measure on simple synthetic weighted networks before applying it to biological signal transduction, revealing insights into the robustness of kinase signalling. Hence, we have considered a framework which can elucidate both global properties of network rewiring and local information on perturbed genes and pathways. We note that whilst useful for understanding information transfer on a network, there are limitations to NTE, most notably it is computationally expensive, even with estimation by simulation. Asides from application of the measure to other phenotype comparisons of network dynamics and refinement of the algorithms for NTE estimation, there is scope for further investigation of our NTE framework along theoretical lines.

Further theoretical questions may consider which ISDs are maintained under different networks, these represent persistent (attractor) states of the network information distribution. To identify such states we note that every graph (if we permit self edges at every vertex to represent a non-zero probability of signal maintenance) admits a disjoint vertex cycle decomposition [261]. Thus there is always a way of sending signal around the network, without combining signal from two vertices at any one vertex. This implies that for every weighted network $\mathcal{G}$, with self edges, there must exist a permutation matrix ϕ, which admits at least one vector $\vec{x}$ satisfying $\phi\vec{x} = \vec{x}$ (*e.g.,* the vector $(1, \ldots, 1)^T$), such that $P[X_1 = \phi\vec{x}|\vec{X_0} = \vec{x}] > 0$. The state $\vec{x}$ thus has a non-zero probability of being a persistent information state of the network.

Questions concerning the evolution of self-assembling networks can also be considered in our framework via an application of dynamic programming. To see this we consider the space $\mathcal{A}$ (constructed above) containing matrix representations of all possible single path length signal forwarding choices, induced by the complete graph on K vertices. We note that for every weighted network, $\mathcal{G}$, on N vertices, where $N < K$, the corresponding stochastic matrix P can be expressed as a convex combination of elements in $\mathcal{A}$, $P = \sum_{j=1}^{K^N} \rho_j A_j$ where $\{A_j\}_{j=1}^{K^N} = \mathcal{A}$, $\sum_j \rho_j = 1$, $\rho_j \leq 1$ for all j. If one interprets the space $\mathcal{A}$ as a state space of possible choices of signal dynamics through the network and considers $\vec{\rho} = \{\rho_j\}_{j=1}^{K^N}$ as a policy, giving a probability distribution of selecting a given global signal dynamic from the state space, that has been obtained by some optimality criterion, then one has a dynamic programming framework for network dynamic evolution. We note that one can calculate the policy selected by a given weighted network explicitly, as $\rho_i = P[\vec{X_1} = A_i\vec{X_0}|\vec{X_0}]$. Thus we have information to guide to construction of an optimality criterion describing network evolution. Forms of such a criterion can be posited and parametrised for different systems and suitable parameter regimes can be reverse engineered from the policy solution $\vec{\rho}$.

2.3 Signalling Entropy: Theoretical Investigations

2.3.1 Introduction

In the previous section we introduced and robustly examined a theoretical framework for traffic on a weighted network. We also introduced and explored NTE, demonstrating its applicability to analysing biological signal transduction. Whilst useful, there are limitations to NTE, however, and it is worth examining other entropy based network theoretic tools derived from our model of network traffic. In this section we thus consider the entropy rate of the Markov chain described by our data derived stochastic matrix model of network traffic. We will focus on the single sample derived stochastic matrix as this is more analytically tractable and has a history of use in a biological context [262]. To avoid ambiguity, we will refer to the entropy rate of such a single sample stochastic matrix as the 'signalling entropy' of the sample.

Formally, we consider a random walk, on an undirected graph $\mathcal{G} = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices and $E = \{(i,j)|i,j \in V\}$ a set of edges, with adjacency matrix $A = (a_{ij})_{ij \in V}$. To each vertex $i \in V$ we assign a variable $x_i \in \mathbb{R}^{>0}$, and denote the vector containing all such variables by $x = (x_i)_{i=1}^n \in \Omega \subset (\mathbb{R}^{>0})^n$, where Ω is a bounded domain, for which there exists $\phi, \omega \in \mathbb{R}^{>0}$ such that for $x \in \Omega$, $x_k \in [\phi, \omega], \forall k \in V$.

We consider a Markov Chain on $\mathcal{G}$ with transition probability matrix $P(x) = (p_{ij}(x))_{ij \in V}$ defined via

$$p_{ij}(x) = \frac{a_{ij}x_j}{\sum_{k \in V} a_{ik}x_k}$$

and define the following measures

1. The degree of vertex $i \in V$ in the interactome, defined by

$$\deg(i) := \sum_k a_{ik}. \tag{26}$$

2. The local entropy of vertex $i \in V$, defined by

$$S_i(x) := -\sum_{j \in V} p_{ij}(x) \log p_{ij}(x). \tag{27}$$

3. The entropy rate of $P(x)$, defined by

$$SR(x) := \sum_{i \in V} \mu_i(x) S_i(x). \tag{28}$$

where μ_i denotes the stationary distribution of $P(x)$ and satisfies

$$\mu_j(x) = \sum_{i \in V} p_{ij}(x)\mu_i(x). \tag{29}$$

We refer to $SR(x)$ as the signalling entropy of x.

We note that we will only be considering this single sample construction, where $\mathcal{G}$ represents a PIN describing proteins which correspond to genes, and x represents a vector of gene or protein expression. Hence to emphasise biological relevance, we will interchangeably use protein/gene to refer to a vertex of $\mathcal{G}$, and use the expression of gene/protein i to refer to x_i.

Theoretical investigations of the entropy rate of a random walk on a complex network have been considered extensively previously. The concept was introduced in [249] as a measure of the minimal amount of information required to describe a diffusion process on a complex network. The measure was considered for degree biased random walks on scale free and real world networks, and was shown to diverge for synthetic scale free networks as the degree exponent tends to 2. Interestingly, it was demonstrated that different real world networks displayed either a consistently higher or lower entropy rate than their synthetic counterparts, with social networks and internet router networks displaying higher entropy rates and airport and hyperlink networks lower. This difference was attributed to degree-degree correlations.

Previous work from our group focused on the idea that oncogenic signal transduction was more disordered compared to its healthy counterpart [23, 17]. This work considered the phenotype level stochastic matrix

$$p_{ij}^{pheno}(x) = \frac{a_{ij}(1 + cor(X_i, X_j))}{\sum_{k \in V} a_{ik}(1 + cor(X_i, X_k))},$$

and demonstrated that 'flux entropy', defined as a weighted sum of local entropies

$$S := \sum_i \frac{1}{deg(i)} S_i(x),$$

was elevated in the cancerous as opposed to healthy phenotype [23] and in metastatic as opposed to non-metastatic breast cancer [17]. It was subsequently demonstrated on the single sample level, using the stochastic matrix described above, that cancerous samples displayed a significantly higher signalling entropy than healthy samples [262].

These results are encouraging and merit further investigation. In particular we wish to attribute clearer biological relevance to the finding that signalling entropy is elevated in cancerous samples, so that the result may find a more solid basis for experimental and clinical investigation. In this section we thus consider theoretical properties of signalling entropy, which may predispose our measure to be an indicator of oncogenic status.

We first demonstrate that the expression of high degree vertices contribute the most to changes in signalling entropy, a finding we validate using proteomic and transcriptomic data describing cancerous tissue. This result suggests that signalling entropy may be detecting the mutations in, and over-expression of, hub genes which drive oncogenesis.

Moreover, the result indicates that signalling entropy may be most easily controlled experimentally by modification of the expression of hub genes.
We subsequently consider intra-sample heterogeneity, and derive a sufficient condition for a sample consisting of two mixed cell types to, on average, have a higher signalling entropy than a homogeneous sample. We validate that this condition holds on transcriptomic data describing 33 distinct tissue types corresponding to 528 pairwise mixtures. This second finding suggests that signalling entropy may be detecting the increased cell type heterogeneity found in cancerous tissue. As such heterogeneity varies between patients and is considered a difficult to measure prognostic factor, this finding motivates the use of our measure as a correlate of clinical outcome, independent of other current measures.
We close with a discussion of further biological implications of these theoretical results. Notably we mention a possible use for signalling entropy as a quantifier of the differentiation potential of a sample, and motivate the data driven investigations of subsequent chapters.

2.3.2 Signalling entropy and high degree vertices

2.3.2.1 Motivation Previous work has demonstrated that signalling entropy is a quantifier of malignant status [262]. It has also been observed that genes which are differentially expressed and mutated in cancer appear to correspond to proteins of higher degree in the PIN [2]. Given this association we wished to investigate whether the differential expression of high degree vertices could be driving the association between signalling entropy and oncogenic status.
We hypothesised that the sensitivity of signalling entropy to changes in the expression of a given gene, may be bounded by an increasing function of the degree of the corresponding protein in the PIN. This hypothesis amounts to the following proposition:

Proposition 1 *Let $x \in \Omega$ and $f : \mathbb{N} \to \mathbb{R}^+$ be a strictly increasing function, then*

$$|\partial_k SR(x)| \leq f(\deg(k)). \tag{30}$$

Where $\partial_k := \frac{\partial}{\partial x_k}$.

Clearly, if the above proposition is correct, then changes in the expression of genes corresponding to high degree proteins, will have greater potential to alter the signalling entropy of a cellular sample. Whence our hypothesis that the differential expression of high degree vertices is a potential driver of signalling entropy's association with malignancy is theoretically sound.

In what follows we prove that this proposition is indeed correct and that the function f can be expressed simply as a multiple of $\deg(k)$.
Subsequently we consider the METABRIC transcriptomic data set of 1980 breast cancers [157] and proteomic data from the Human Protein Atlas [263] corresponding to 20 cancerous tissue types to provide empirical validation of our hypothesis. We demonstrate that genes whose expression variations are most correlated with signalling entropy variations are more likely to be of higher degree. These results provide strong evidence that the association of signalling entropy with malignancy is likely related to the differential expression of genes corresponding to highly connected proteins, and suggest the targeting of such proteins as a strategy to control signalling entropy experimentally.

2.3.2.2 Proof of Proposition 1 In this section we prove the following version of Proposition 1

Theorem 2 *Let $x \in \Omega$, such that $x_k \in [\phi, \omega], \forall k \in V$, then*

$$|\partial_k SR(x)| \leq \frac{\omega \left| 2SR(x) - \log \frac{\omega^2 |V|^2}{\phi^2} \right|}{W(x)} \deg(k). \tag{31}$$

Proof: We first note that the a closed form expression for signalling entropy can easily be derived as described in [264]. To demonstrate this let us define the following quantities

$$W_{ij}(x) := a_{ij} x_i x_j$$

$$W_i(x) := \sum_j W_{ij}(x)$$

$$W(x) := \sum_i W_i(x).$$

By multiplying $p_{ij}(x)$ by $1 = x_i/x_i$, we see that

$$p_{ij} = \frac{W_{ij}(x)}{W_i(x)}.$$

Thus $P(x)$ describes a weighted random walk on an undirected graph and it thus follows that

$$\mu_i(x) = W_i(x)/W(x). \tag{32}$$

The proof of (32) follows from simple substitution into (29):

$$\begin{aligned} LHS &= \sum_i p_{ij}(x)\mu_i(x) \\ &= \sum_i \frac{W_{ij}(x)}{W_i(x)} \frac{W_i(x)}{W(x)} \\ &= \frac{W_j(x)}{W(x)} \\ &= RHS. \end{aligned}$$

From this result it follows that

$$SR(x) = -\frac{1}{W(x)} \sum_{ij} W_{ij}(x) \log \frac{W_{ij}(x)}{W_i(x)}. \tag{33}$$

By differentiation of (33) with respect to x_k we obtain

$$\begin{aligned} \partial_k SR(x) &= \partial_k \left(-\frac{1}{W(x)} \sum_{ij} W_{ij}(x) \log \frac{W_{ij}(x)}{W_i(x)} \right) \\ &= \frac{\partial_k W(x)}{W(x)^2} \sum_{ij} W_{ij}(x) \log \frac{W_{ij}(x)}{W_i(x)} - \\ &\quad \frac{1}{W(x)} \sum_{ij} \left[\partial_k W_{ij}(x) \log \frac{W_{ij}(x)}{W_i(x)} + W_{ij}(x) \left(\frac{\partial_k W_{ij}(x)}{W_{ij}(x)} - \frac{\partial_k W_i(x)}{W_i(x)} \right) \right] \\ &= -\frac{\partial_k W(x)}{W(x)} SR(x) - \frac{1}{W(x)} \sum_{ij} \left[\partial_k W_{ij} \left(1 + \log \frac{W_{ij}(x)}{W_i(x)} \right) - \right. \\ &\quad \left. \partial_k W_i(x) \frac{W_{ij}(x)}{W_i(x)} \right]. \end{aligned} \tag{34}$$

It is sensible to next evaluate the partial derivatives of $W_{ij}(x), W_i(x)$ and $W(x)$.

$$\begin{aligned}
\partial_k W_{ij}(x) &= \partial_k(a_{ij}x_i x_j) \\
&= a_{ij}(x_j\delta_{ik} + x_i\delta_{jk}). \qquad (35)\\
\partial_k W_i(x) &= \sum_j \partial_k W_{ij}(x) \\
&= \sum_j a_{ij}(x_j\delta_{ik} + x_i\delta_{jk}) \\
&= x_i a_{ik} + \delta_{ik}\sum_j a_{ij}x_j. \qquad (36)\\
\partial_k W(x) &= \sum_i \partial_k W_i(x) \\
&= \sum_i \left(x_i a_{ik} + \delta_{ik}\sum_j a_{ij}x_j \right) \\
&= \sum_i x_i a_{ik} + \sum_j a_{kj}x_j \\
&= 2\sum_i x_i a_{ik}. \qquad (37)
\end{aligned}$$

Where the last line was obtained by appealing to the undirected nature of the interactome which ensures $a_{ij} = a_{ji}$. Substituting these expressions into (34) we obtain:

$$\begin{aligned}
\partial_k SR(x) &= -\frac{2\sum_i x_i a_{ik}}{W(x)} SR(x) - \\
&\frac{1}{W(x)}\sum_{ij}\left[a_{ij}(x_j\delta_{ik} + x_i\delta_{jk})\left(1 + \log\frac{W_{ij}(x)}{W_i(x)}\right) - \right. \\
&\left. \left(x_i a_{ik} + \delta_{ik}\sum_l a_{il}x_l \right)\frac{W_{ij}(x)}{W_i(x)} \right]. \qquad (38)
\end{aligned}$$

Considering just the second term in the above expression we see

$$\begin{aligned}
&-\frac{1}{W(x)}\sum_{ij}\left[a_{ij}(x_j\delta_{ik} + x_i\delta_{jk})\left(1 + \log\frac{W_{ij}(x)}{W_i(x)}\right) - \left(x_i a_{ik} + \delta_{ik}\sum_l a_{il}x_l \right)\frac{W_{ij}(x)}{W_i(x)} \right] \\
= -\frac{1}{W(x)}&\left[\sum_j a_{kj}x_j\left(1 + \log\frac{W_{kj}(x)}{W_k(x)}\right) + \sum_i a_{ik}x_i\left(1 + \log\frac{W_{ik}(x)}{W_i(x)}\right) - \sum_{ij} x_i a_{ik}\frac{W_{ij}(x)}{W_i(x)} \right. \\
&\left. - \sum_{jl} a_{kl}x_l\frac{W_{kj}(x)}{W_k(x)} \right]. \qquad (39)
\end{aligned}$$

We again appeal to $a_{ij} = a_{ji}$, which implies that $W_{ij}(x) = W_{ji}(x)$, to obtain:

$$\begin{aligned}
&-\frac{1}{W(x)}\left[\sum_j a_{kj}x_j\left(1+\log\frac{W_{kj}(x)}{W_k(x)}\right)+\sum_i a_{ik}x_i\left(1+\log\frac{W_{ik}(x)}{W_i(x)}\right)\right.\\
&\qquad\left.-\sum_{ij}x_i a_{ik}\frac{W_{ij}(x)}{W_i(x)}-\sum_{jl}a_{kl}x_l\frac{W_{kj}(x)}{W_k(x)}\right]\\
&=-\frac{1}{W(x)}\left[\sum_i x_i a_{ik}\left(2+\log\frac{W_{ik}(x)^2}{W_k(x)W_i(x)}\right)-\sum_{ij}x_i a_{ik}\left(\frac{W_{ij}(x)}{W_i(x)}+\frac{W_{kj}(x)}{W_k}\right)\right]\\
&=-\frac{1}{W(x)}\left[\sum_i x_i a_{ik}\left(2+\log\frac{W_{ik}(x)^2}{W_k(x)W_i(x)}\right)-\sum_i x_i a_{ik}\left(\frac{\sum_j W_{ij}(x)}{W_i(x)}+\frac{\sum_j W_{kj}(x)}{W_k(x)}\right)\right]\\
&=-\frac{1}{W(x)}\sum_i x_i a_{ik}\log\frac{W_{ik}(x)^2}{W_k(x)W_i(x)}.
\end{aligned} \tag{40}$$

Substitution of this term into (38) we obtain

$$\partial_k SR(x) = -\frac{1}{W(x)}\sum_i x_i a_{ik}\left(2SR(x)+\log\frac{W_{ik}(x)^2}{W_k(x)W_i(x)}\right). \tag{41}$$

Given that

$$\frac{W_{ik}(x)^2}{W_k(x)W_i(x)} \geq \frac{\phi^2}{\omega^2|V|^2}, \tag{42}$$

it is a simple deduction that

$$|\partial_k SR(x)| \leq \frac{\omega\left|2SR(x)-\log\frac{\omega^2|V|^2}{\phi^2}\right|}{W(x)}\deg(k). \tag{43}$$

Hence the theorem is proven.

□

It has been demonstrated that the maximal entropy rate of a random walk on an adjacency matrix $(a_{ij})_{ij\in V}$, with largest eigen value λ and corresponding eigen vector $\nu = (\nu_i)_{i\in V}$ is $\log\lambda$ and is achieved when $p_{ij} = \frac{a_{ij}\nu_j}{\lambda\nu_i}$ [274]. We note that this corresponds in our framework to $SR(\nu) = \log\lambda$. It is easily proven by substitution that (41) evaluates to zero when $x = \nu$, in line the maximal entropy rate being achieved when the gene expression profile matches the eigen vector centrality of the PIN.

2.3.2.3 Empirical Validation In the above section we demonstrated that the sensitivity of signalling entropy to changes in the expression of a gene is bounded by a

multiple of the degree of the corresponding protein. This result suggests that expression differences between cancerous and healthy tissue, which are dominantly in genes corresponding to high degree proteins, strongly influence our measure. Here we further validate this hypothesis by consideration of the METABRIC transcriptomic data set of 1980 breast cancers divided into discovery and validation data sets of equal proportion, and the Human Protein Atlas dataset describing proteomic data for 20 healthy and cancerous tissues.

To ascertain which gene expression patterns are most associated with signalling entropy, we first computed the signalling entropy of each sample in both the discovery and validation data sets of METABRIC and for the cancerous samples in the Human Protein Atlas (Materials and Methods, Chapter 2). We then evaluated the Pearson correlation between gene expression and signalling entropy for each gene corresponding to a protein in our PIN, across all samples in the discovery and validation data sets of METABRIC and all cancers in the Human Protein Atlas separately.

As expected, we found that there was a highly significant correlation between the a gene's association with signalling entropy, and its corresponding degree in the PIN ($p < 2.2 \times 10^{-16}$, Fig 2.5). This confirms our theoretical result that higher degree proteins are more associated to changes in (and hence are more correlated with) signalling entropy.

2.3.2.4 Summary In this section we have demonstrated that the sensitivity of signalling entropy to changes in the expression of a given gene is bounded by a multiple of the degree of the corresponding protein in the PIN. Whence it is to be anticipated that expression changes in genes corresponding to high degree proteins are more likely to drive changes in signalling entropy. By considering transcriptomic and proteomic data from primary tumour samples, we empirically confirmed this postulate, demonstrating that genes whose expression was most strongly correlated with signalling entropy across tumour samples were more likely to be of high degree.

The results in this section thus make the case that elevated signalling entropy in cancerous tissue, may be driven by mutation or over-expression of high degree genes and marks out such genes as the organisers of order in interactome signalling. Experimental and therapeutic studies aimed at reducing signalling entropy in cancer may thus benefit from a focus on the modification of expression of high degree proteins.

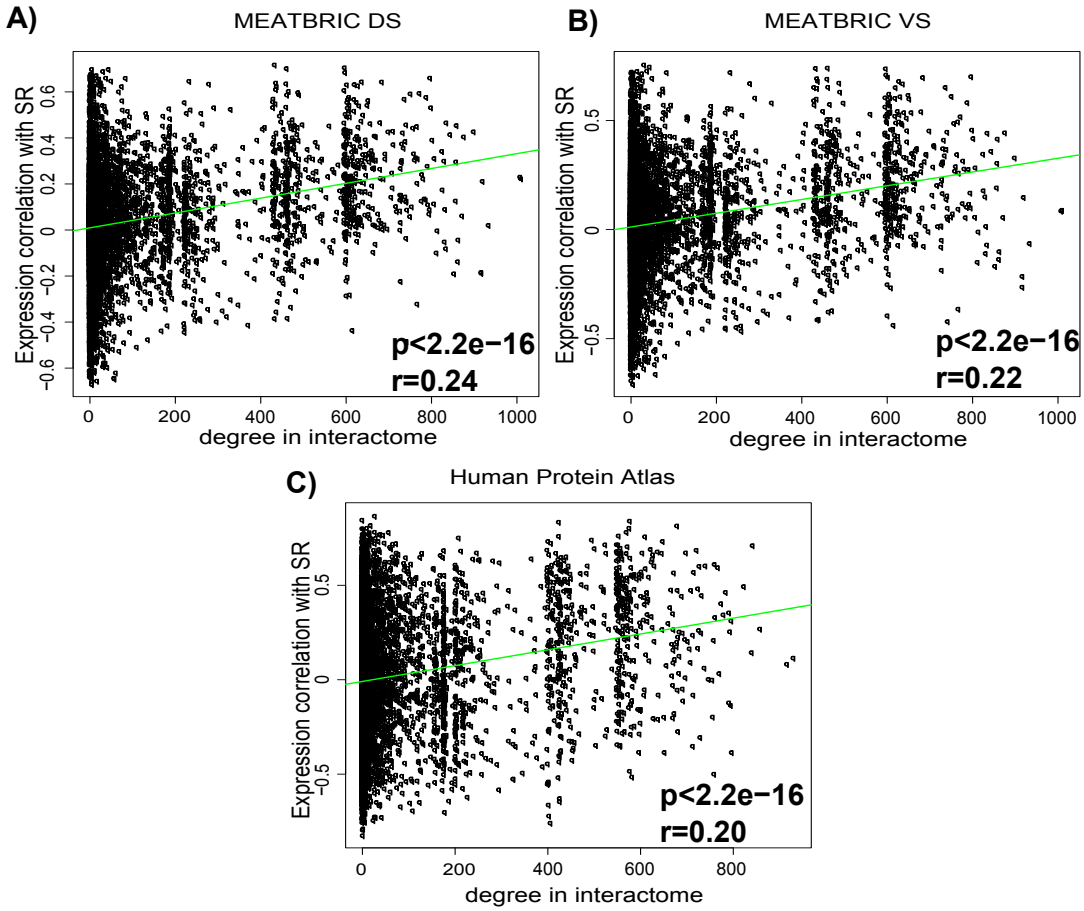

Figure 2.5: **Signalling entropy is driven by the expression of high degree vertices.** Correlation between degree in the PIN and Pearson correlation between gene expression and signalling entropy in the MEATABRIC discovery dataset (A), the METABRIC validation dataset (B) and the Human Protein Atlas cancer samples (C). We see that high degree vertices have gene expression variances more closely associated to signalling entropy.

2.3.3 Signalling Entropy and heterogeneity

2.3.3.1 Motivation Intra-sample heterogeneity, the diversity of cell types within a cellular sample has long been observed as higher in tumours than healthy tissue, and we therefore postulated that an elevated signalling entropy in cancerous tissue may be driven by a more diverse cell population.

To investigate whether signalling entropy does associate with intra-sample heterogeneity, we consider our measure evaluated for three theoretical samples: namely two homogeneous samples consisting only of cell type x or y respectively, and a third heterogeneous sample consisting of a 50:50 mixture of cell types x and y. It is clear that if cell type x has an expression profile that maximises signalling entropy and cell type y does not, then the signalling entropy of the mixture will be lower than the signalling entropy of x, thus signalling entropy is not a point-wise measure of heterogeneity. However, most biologically realistic cell types have distinct expression profiles, corresponding to the existence of non-overlapping active pathways between cell type pairs [265]. We therefore posited that the signalling entropy of a mixed sample may be higher than that of a homogeneous sample on average.

In what follows, we first show that it is a consequence of simple algebra that if signalling entropy is super-additive over the set of biologically admissible expression profiles (*i.e.,* Signalling Entropy$\left(\frac{x+y}{2}\right) > \frac{1}{2}$ Signalling Entropy$(x) + \frac{1}{2}$ Signalling Entropy(y)) then signalling entropy will on average be elevated in mixed samples as opposed to homogeneous samples. We thus derive a condition for point-wise super-additivity of our measure, which we show is typically valid for biologically realistic data sets. Following the derivation of this condition we consider data corresponding to gene expression profiles for 33 distinct adult tissues, representing 528 possible pairwise homogeneous mixtures [265]. We demonstrate that our condition for super-additivity and hence for elevated signalling entropy in heterogeneous samples holds for all 528 mixtures, before demonstrating that signalling entropy is indeed elevated in the mixed samples on average compared to the homogeneous tissue samples.

These results provide strong evidence that signalling entropy associates with intra-sample heterogeneity on average. Hence elevated signalling entropy in cancer may be a consequence of increased heterogeneity in cancerous tissue.

2.3.3.2 Super-additivity implies signalling entropy is elevated in a mixed sample on average We hypothesised that the signalling entropy of a heterogeneous sample generated from a 50:50 mixture of two homogeneous cell types will be greater, on average, than the signalling entropy of a homogeneous sample. Here we show that if

signalling entropy is super-additive then the hypothesis is correct. Let us first define some preliminaries: Let $x_i \in \mathbb{R}^{>0}$ be the expression of gene i in cell type X, and denote the vector containing all such variables by $x = (x_i)_{i=1}^n \in \Omega \subset \mathbb{R}^{>0}$, where Ω is some bounded domain. In our analysis x will represent the vector of normalised gene expression values for a homogeneous sample.

Our hypothesis that signalling entropy is elevated in heterogeneous samples on average amounts to proving the following proposition:

Proposition 2 *Let $x, y \in \Omega$, then*

$$\int_\Omega \int_\Omega \left(SR\left(\frac{x+y}{2}\right) - SR(x) \right) dxdy > 0. \tag{44}$$

Let us consider the following claim:

Claim 3 (Super-additivity) *Let $x, y \in \Omega$ then*

$$SR\left(\frac{x+y}{2}\right) > \frac{SR(x)}{2} + \frac{SR(y)}{2}. \tag{45}$$

It is clear that if the claim is true then the proposition must be true. Notice first that if the claim is true then as it is a strict bound $\exists \epsilon > 0$ such that $SR\left(\frac{x+y}{2}\right) > \frac{SR(x)}{2} + \frac{SR(y)}{2} + \epsilon$. Whence

$$\begin{aligned} \int_\Omega \int_\Omega \left(SR\left(\frac{x+y}{2}\right) - SR(x) \right) dxdy &> \int_\Omega \int_\Omega \left(\frac{SR(y)}{2} - \frac{SR(x)}{2} + \epsilon \right) dxdy & (46) \\ &= |\Omega|^2 \epsilon & (47) \\ &> 0, & (48) \end{aligned}$$

and thus the proposition is true.

Thus if signalling entropy is super-additive over homogeneous cell types, this implies that signalling entropy will on average be elevated in heterogeneous mixtures of cell types.

2.3.3.3 A sufficient condition for signalling entropy to be super-additive

Here we will prove the following theorem

Theorem 4 *Let $x, y \in \Omega$, let $a = \max_i \left(\frac{x_i}{y_i}\right)$, and let $b = \min_i \left(\frac{x_i}{y_i}\right)$. A sufficient condition for Claim 3 to be true is*

$$sign(1 - 1/b + 2/a) + sign(1 - a + 2b) = 2 \tag{49}$$

Proof: From (33)

$$SR((x+y)/2) = -\frac{1}{W((x+y)/2)} \sum_{ij} W_{ij}((x+y)/2) \log \frac{W_{ij}((x+y)/2)}{W_i((x+y)/2)}, \tag{50}$$

it is prudent to first consider $W_{ij}((x+y)/2)$:

$$W_{ij}((x+y)/2) = \frac{a_{ij}}{4}(x_i+y_i)(x_j+y_j) \quad (51)$$

$$= \frac{1}{4}(W_{ij}(x)+W_{ij}(y)+a_{ij}(x_iy_j+x_jy_i)). \quad (52)$$

We will define

$$\hat{W}_{ij}(x,y) = a_{ij}(x_iy_j+x_jy_i) \quad (53)$$

$$\hat{W}_i(x,y) = \sum_j \hat{W}_{ij}(x,y) \quad (54)$$

$$\hat{W}(x,y) = \sum_j \hat{W}_i(x,y) \quad (55)$$

for notational ease. Note that as $y, x > 0$

$$\hat{W}_{ij}(x,y) = \frac{x_i}{y_i}W_{ij}(y) + \frac{y_i}{x_i}W_{ij}(x). \quad (56)$$

$$\hat{W}_i(x,y) = \frac{x_i}{y_i}W_i(y) + \frac{y_i}{x_i}W_i(x). \quad (57)$$

It thus follows that:

$$SR((x+y)/2) = \frac{-1}{W(x)+W(y)+\hat{W}(x,y)}\sum_{ij}(W_{ij}(x)+W_{ij}(y)+ \hat{W}_{ij}(x,y))\log\frac{W_{ij}(x)+W_{ij}(y)+\hat{W}_{ij}(x,y)}{W_i(x)+W_i(y)+\hat{W}_i(x,y)}. \quad (58)$$

We will now appeal to the log sum inequality:

Theorem 5 (Log-sum inequality) *Let $a_1, ..., a_n$, $b_1, ..., b_n$ be non-negative numbers then*

$$\sum_{i=1}^{n} a_i \log\frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right)\log\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}. \quad (59)$$

If we denote $a_1 = W_{ij}(x), a_2 = W_{ij}(y), a_3 = \hat{W}_{ij}(x,y)$ and $b_1 = W_i(x), b_2 = W_i(y), b_3 = \hat{W}_i(x,y)$, and apply the log-sum inequality to the summand of (58) we obtain:

$$SR((x+y)/2) \geq \frac{-1}{W(x)+W(y)+\hat{W}(x,y)}\sum_{ij} W_{ij}(x)\log\frac{W_{ij}(x)}{W_i(x)} + \quad (60)$$

$$W_{ij}(y)\log\frac{W_{ij}(y)}{W_i(y)} + \hat{W}_{ij}(x,y)\log\frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)} \quad (61)$$

$$= \frac{W(x)SR(x)+W(y)SR(y)-\sum_{ij}\hat{W}_{ij}(x,y)\log\frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)}}{W(x)+W(y)+\hat{W}(x,y)}. \quad (62)$$

Now consider the term $-\sum_{ij}\hat{W}_{ij}(x,y)\log\frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)}$ and apply the log sum inequality again:

$$\begin{aligned}-\sum_{ij}\hat{W}_{ij}(x,y)\log\frac{\hat{W}_{ij}(x,y)}{\hat{W}_i(x,y)} &= -\sum_{ij}\left(\frac{x_i}{y_i}W_{ij}(y)+\frac{y_i}{x_i}W_{ij}(x)\right)\log\frac{\frac{x_i}{y_i}W_{ij}(y)+\frac{y_i}{x_i}W_{ij}(x)}{\frac{x_i}{y_i}W_i(y)+\frac{y_i}{x_i}W_i(x)}\\ &\geq -\sum_{ij}\frac{x_i}{y_i}W_{ij}(y)\log\frac{W_{ij}(y)}{W_i(y)}+\frac{y_i}{x_i}W_{ij}(x)\log\frac{W_{ij}(x)}{W_i(x)}. \quad (63)\end{aligned}$$

Returning to (60), it now follows that

$$\begin{aligned}SR\left(\frac{x+y}{2}\right)-\frac{SR(x)-SR(y)}{2} \geq\\ \frac{1}{2(W(x)+W(y)+\hat{W}(x,y))}\Bigg(SR(x)(W(x)-W(y)-\hat{W}(xy))+\\ SR(x)(W(y)-W(x)-\hat{W}(xy))-\\ 2\sum_{ij}\frac{x_i}{y_i}W_{ij}(y)\log\frac{W_{ij}(y)}{W_i(y)}+\frac{y_i}{x_i}W_{ij}(x)\log\frac{W_{ij}(x)}{W_i(x)}\Bigg). \quad (64)\end{aligned}$$

So Claim 3 is true if the LHS of the above equation is positive, *i.e.*,

$$\begin{aligned}0<\frac{1}{2(W(x)+W(y)+\hat{W}(x,y))}\Bigg(SR(x)(W(x)-W(y)-\hat{W}(xy))+\\ SR(x)(W(y)-W(x)-\hat{W}(xy))-\\ 2\sum_{ij}\frac{x_i}{y_i}W_{ij}(y)\log\frac{W_{ij}(y)}{W_i(y)}+\frac{y_i}{x_i}W_{ij}(x)\log\frac{W_{ij}(x)}{W_i(x)}\Bigg). \quad (65)\end{aligned}$$

We note that if (65) holds then

$$\begin{aligned}2\left(W(x)SR(x)+W(y)SR(y)-\sum_{ij}\frac{x_i}{y_i}W_{ij}(y)\log\frac{W_{ij}(y)}{W_i(y)}+\frac{y_i}{x_i}W_{ij}(x)\log\frac{W_{ij}(x)}{W_i(x)}\right)>\\ (SR(x)+SR(y))(W(x)+W(y)+\hat{W}(xy)) \quad (66)\end{aligned}$$

We note that the RHS of (66) satisfies

$$\begin{aligned}2(W(x)SR(x)+W(y)SR(y)-\sum_{ij}\frac{x_i}{y_i}W_{ij}(y)\log\frac{W_{ij}(y)}{W_i(y)}+\frac{y_i}{x_i}W_{ij}(x)\log\frac{W_{ij}(x)}{W_i(x)}\\ >2W(x)SR(x)(1+\min_i\frac{y_i}{x_i})+2W(y)SR(y)(1+\min_i\frac{x_i}{y_i}). \quad (67)\end{aligned}$$

We further notice that

$$\begin{aligned}\hat{W}(xy) &= \sum_i\frac{x_i}{y_i}W_i(y)+\frac{y_i}{x_i}W_i(x) \quad (68)\\ &< \max_i\frac{x_i}{y_i}W(y)+\max_i\frac{y_i}{x_i}W(x) \quad (69)\end{aligned}$$

whence the LHS of (66) satisfies

$$(SR(x)+SR(y))(W(x)+W(y)+\hat{W}(xy)) < SR(x)W(x)(1+\max_i \frac{y_i}{x_i})+SR(y)W(y)(1+\max_i \frac{x_i}{y_i}). \quad (70)$$

Thus it follows that if

$$SR(x)W(x)(1+\max_i \frac{y_i}{x_i})+SR(y)W(y)(1+\max_i \frac{x_i}{y_i}) < 2W(x)SR(x)(1+\min_i \frac{y_i}{x_i})+2W(y)SR(y)(1+\min_i \frac{x_i}{y_i}), \quad (71)$$

and consequentially

$$SR(x)W(x)(1-\max_i \frac{y_i}{x_i}+2\min_i \frac{y_i}{x_i})+SR(y)W(y)(1-\max_i \frac{x_i}{y_i}+2\min_i \frac{x_i}{y_i})>0$$

$$SR(x)W(x)(1-1/b+2/a)+SR(y)W(y)(1-a+2b)>0 \quad (72)$$

then LHS<RHS. Where $a=\max_i\left(\frac{x_i}{y_i}\right)$, and $b=\min_i\left(\frac{x_i}{y_i}\right)$, as in the claim. We note that as $SR(x)W(x)>0$ and $SR(y)W(y)>0$ that the condition will always hold provided

$$sign(1-1/b+2/a)+sign(1-a+2b)=2.$$

Hence the theorem is correct.

□

We note that the condition can be computed numerically for a range of values and is valid for the majority of biologically plausible ranges, with the exception of samples with extremely low expression value ratios (Fig 2.6).

2.3.3.4 Empirical validation of super-additivity of signalling entropy on Ω If the bound derived above holds in Ω, the space of biologically admissible homogeneous sample expression regimes, then Claim 3 and hence Proposition 2 are true and we have proven our postulate that signalling entropy is elevated on average in mixed biological samples.

Though we have shown that most biologically realistic expression values adhere to this bound, for very low maximum expression ratio values the bound can fail. Such low expression ratio values are unlikely to be biologically realistic, however for rigour we here consider the validity of the bound in a biological context, via a data set described in GSE2361 [265] (Materials and Methods, Chapter 2). The data set profiles 33 distinct adult tissues and 3 foetal tissues. We will disregard the foetal tissues as the tissue types overlap with the adult tissues and thus cannot be considered distinct homogeneous samples.

We first computed the value of $sign(1 - 1/b + 2/a) + sign(1 - a + 2b)$ for every pairwise combination of the 33 tissue types (528 pairwise combinations). We found that for every combination of heterogeneous samples the condition described in Theorem 4 was satisfied and thus Claim 3 (super-additivity of signalling entropy) was correct for this data set (Fig 2.7).

We next computed signalling entropy for the 33 homogeneous tissue types and the 528 pairwise mixtures of samples. As explained above signalling entropy cannot be expected to be a point-wise measure of heterogeneity, and it is certainly true that not every tissue, when mixed with another, displays an increased signalling entropy. We do, however, see a trend towards this happening, with the reproductive tissues (testes, uterus, ovaries, breast) providing the most significant increases (Fig 2.8).

Whence, as expected, Proposition 2 is indeed correct for this data set, and the signalling entropy of the mixed samples is higher than the homogeneous samples on average (paired Wilcoxon test $p = 0.012$, Fig 2.9).

2.3.3.5 Summary In this section we have derived a condition which if satisfied by homogeneous samples guarantees that signalling entropy will be higher in heterogeneous 50:50 mixtures of two homogeneous samples, as compared to the unmixed samples, on average. We then verified that for a large number of homogeneous adult tissues the condition was indeed satisfied and signalling entropy was higher in 50:50 mixtures of homogeneous tissues on average.

These results suggest that signalling entropy represents a quantifier of intra-sample sample heterogeneity, being correlated with heterogeneity at the population level. Hence elevated signalling entropy in cancer may be driven by an increased heterogeneity of tumours as

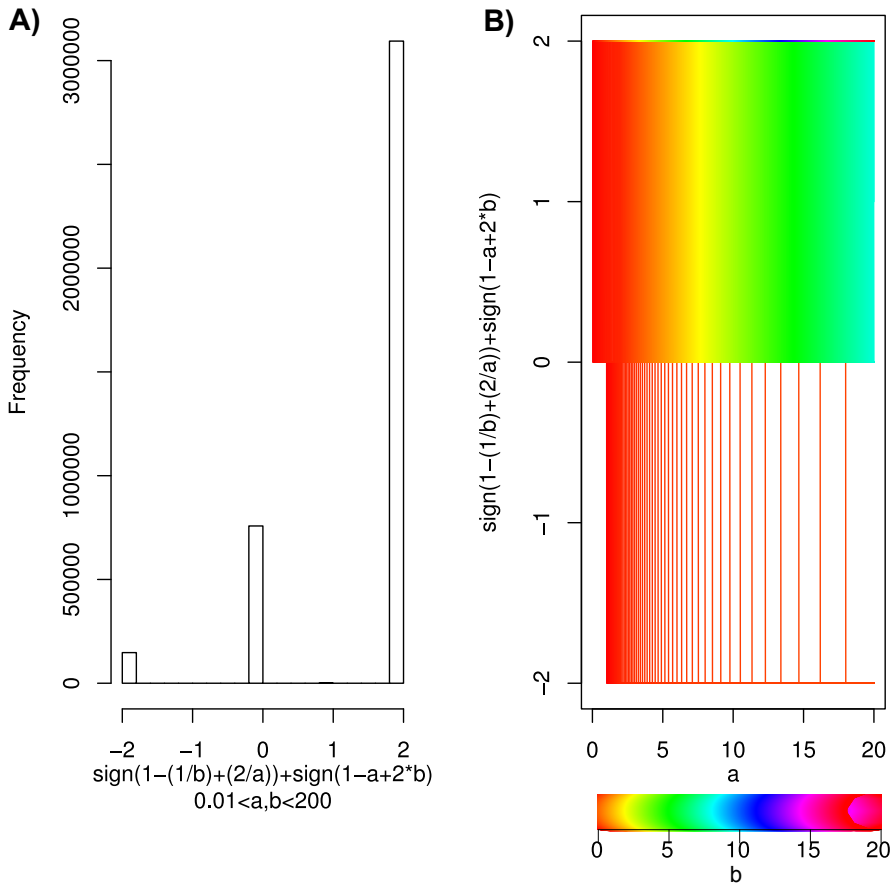

Figure 2.6: **Analysis of the expression** $sign(1-1/b+2/a)+sign(1-a+2b)$ **for a range of biologically plausible values:** $a, b \in [0.01, 20]$**.** (A) Histogram of values of the $sign(1-1/b+2/a)+sign(1-a+2b)$ evaluated over 2000 equally incremented values of a and b over the range $a, b \in [0.01, 20]$. We see that the majority of the values satisfy the condition $sign(1-1/b+2/a)+sign(1-a+2b)=2$. (B) Plot of $sign(1-1/b+2/a)+sign(1-a+2b)$ for $a, b \in [0.01, 20]$, values of a are plotted on the x axis whilst colors from red to green to blue denote increasing values of b, we see that as a and b increase the condition very quickly becomes true, suggesting that the condition only fails for very low maximal expression value ratios, unlikely to be observed in reality.

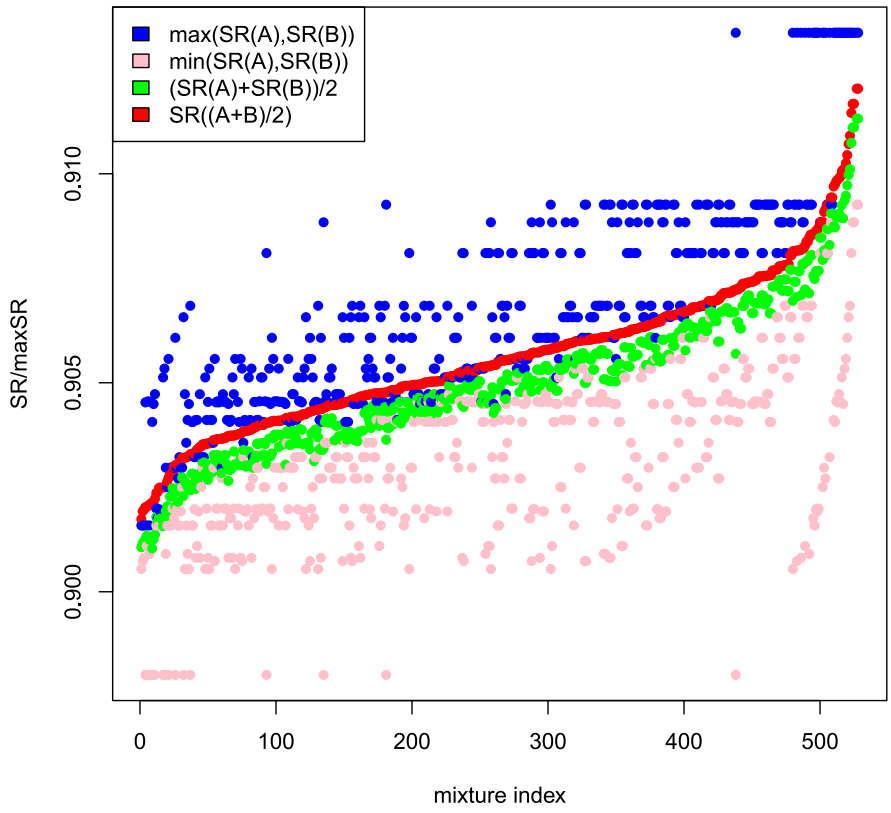

Figure 2.7: **Demonstration that the claim of super-additivity $SR\left(\frac{x+y}{2}\right) > \frac{SR(x)}{2} + \frac{SR(y)}{2}$ is correct for all pairwise combinations of samples in GSE2361.** Signalling entropy (denoted SR/max SR) is computed for 528 distinct pairwise mixtures of 33 homogeneous tissue samples demonstrating that our measure is super-additive and hence will be raised, on average, in mixed samples compared to homogeneous samples.

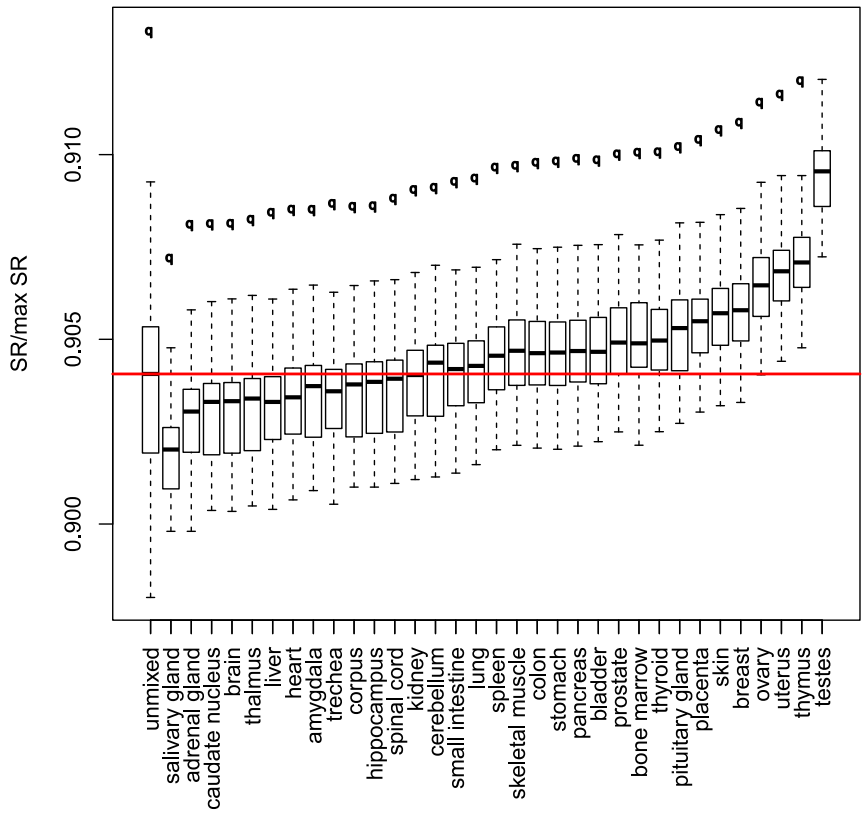

Figure 2.8: **Signalling entropy of homogeneous and mixed tissues.** The first box represents the signalling entropy distribution of unmixed tissues, whilst each subsequent labelled box represents the signalling entropy distribution of the labelled tissue with each of the remaining 32 tissues. The red line represents the median of the unmixed samples. We see that for 20/33 tissue types, the median of the mixture is greater than the median of the pure samples, suggesting that on average the signalling entropy of the mixture is greater than the signalling entropy of the pure sample.

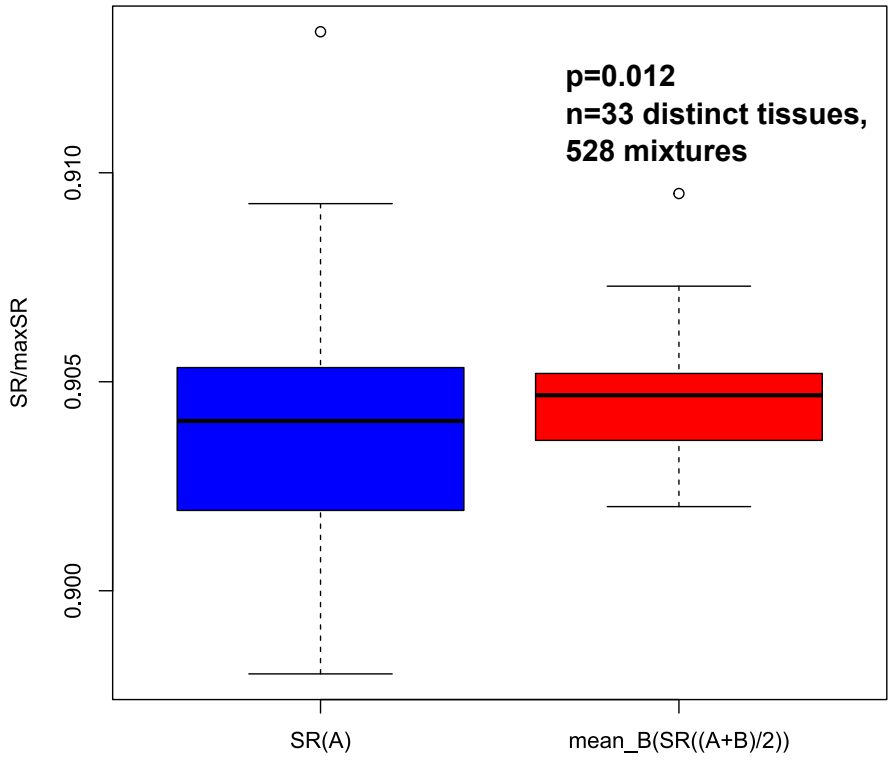

Figure 2.9: **Demonstration that the proposition** $\int_\Omega \int_\Omega \left(SR\left(\frac{x+y}{2}\right) - SR(x)\right) dxdy > 0$ **is correct for samples in GSE2361.** Signalling entropy is raised on average in mixed samples as compared to homogeneous samples, considering 528 mixtures of 33 homogeneous tissue samples. The p-value corresponds to a two tailed paired Wilcoxon signed rank test, and reveals that signalling entropy is significantly elevated in the admixed cell populations on average.

compared to healthy tissue.

2.3.4 Conclusions and Future Directions

Here we have considered signalling entropy, defined as the entropy rate of the Markov chain described by a sample specific stochastic matrix. This measure has been demonstrated as elevated in cancerous tissue and we here investigated two key properties of signalling entropy, which may provide biological insight into this finding, analytically and with empirical validation.

Firstly we proved that the sensitivity of signalling entropy to changes in the expression of a given protein is bounded by a multiple of the degree of that protein in the PIN. Suggesting that signalling entropy may be elevated in cancer as a consequence of over-expression/mutation of hub genes and that such genes may represent targets for the experimental control of signalling entropy.

Secondly we derived and validated a condition for signalling entropy to be raised in heterogeneous 50:50 mixtures of two homogeneous samples, as compared to the unmixed samples on average. Suggesting that signalling entropy may be elevated in cancer, as a consequence of the heterogeneity of tumours compared to healthy tissues.

Both these results thus have clear implications in a cancerous setting. Given that intra-tumour heterogeneity has long been considered a difficult to measure prognostic indicator in oncology [266], the latter result motivates the investigation of signalling entropy in a prognostic context. If signalling entropy is indeed prognostic, lowering it may represent a novel therapeutic strategy. Hence the result that our measure is most strongly influenced by hub genes may become important in therapeutic development.

In the introduction we also discussed employing a network entropy such as signalling entropy as a correlate of a sample's differentiation potential. Given that pluripotent stem cell populations have been characterised by an increased heterogeneity at the population level, the results in this section suggests that our measure is a good candidate quantifier of cell potency.

These theoretical results thus strongly motivate further data driven work on establishing the biological validity of signalling entropy in the context of cellular differentiation and cancer. Whence in Chapter 3, by considering over 1000 genome wide gene expression samples of healthy tissue, we will demonstrate that signalling entropy is elevated stem cells as opposed to differentiated tissue, and deceases systematically during differentiation. We also reveal that signalling entropy is elevated in CSCs as opposed to the tumour bulk. In Chapter 4 we further demonstrate that our measure is strongly prognostic in epithelial cancer, through the consideration of over 5000 primary tumour samples. Finally in Chap-

ter 5, we consider signalling entropy in the context of muscular dystrophy, demonstrating that it is elevated as a consequence of over-expression of the FSHD candidate gene *DUX4*.

2.4 Interactome Sparsification and Rewiring

2.4.1 Introduction

In the previous two subsections of this chapter we have introduced two network theoretic tools: NTE and signalling entropy. Our aim of this chapter, as outlined in the introduction, is the development of entropy based network rewiring methodologies, which can provide local information about gene and pathway rewiring between phenotypes, as well as elucidation of some general, global principles of network rewiring.

We have seen that NTE is powerful for tracing local paths of differential network information flow between two biological phenotypes and that the measure can be derived from a metric space framework which can provide general insights into network rewiring in response to drug perturbation. Thus the NTE measure does indeed satisfy our criteria. However, the algorithm to compute NTE is computationally expensive and does not perform well on large networks.

We have also seen how signalling entropy may represent a global measure of signalling disorder and intra-sample heterogeneity, which can discriminate between cancerous and healthy tissue and may provide insights into cell potency and prognosis in developmental pathology.

Given the global power of signalling entropy and our goal to develop scalable methodologies, we here propose an algorithm based on a local analogue of signalling entropy, which we name *Interactome Sparsification and Rewiring* (InSpiRe).

The algorithm is designed to extract a subset of the human PIN, describing genes and interactions which are significantly rewired between two biological phenotypes described by gene expression data. In contrast to NTE the algorithm is easy to evaluate on large networks.

As the InSpiRe algorithm is intended to extract critical genes and interactions which are rewiring between two biological phenotypes, a full evaluation of its validity requires a biological context. In such a setting pathways which are expected to be detected as a consequence of previous literature can be outlined in detail, and InSpiRe detected pathways which are unexpected can be validated as providing novel insight via experimental models. Such an evaluation is hence a detailed undertaking, requiring experimental biology and thus is not best positioned in this subsection on mathematical analysis of network

theoretic tools. Rather we present here simply an outline of the application of InSpiRe, noting how it can be considered a local analogue of signalling entropy. A full validation of the algorithm in a biological context (that of FSHD), including the experimental validation of novel pathways and the comparison of the algorithm to other methodologies can be found in Chapter 5.

2.4.2 The three steps of the InSpiRe algorithm

InSpiRe is a differential network methodology, consisting of three main steps displayed in Fig 2.10 and described below.

2.4.2.1 Step 1: Integration of mRNA expression data with the PIN The first step of InSpiRe is the integration of expression data with a PIN to create a phenotype level stochastic matrix describing network traffic as explained above (Fig 2.10A). Briefly, for each phenotype we integrate expression data with the PIN by assigning each interaction connecting two proteins, i and j, with a transformed Pearson correlation in the gene expression profiles of proteins i and j across the samples corresponding to the given phenotype r (denoted C^r_{ij}). This results in a weighted network w^r_{ij} which is then transformed into a stochastic matrix p^r_{ij} where

$$p^r_{ij} = \frac{w^r_{ij}}{\sum_{k\in\mathcal{N}(i)} w^r_{ik}}, \tag{73}$$

where $\mathcal{N}(i)$ denotes the set of neighbours of protein i in the PIN.
We interpret row k in the matrix p^r_{ij} as a probability distribution describing the interaction preferences of protein k in phenotype r. Note that $\sum_{j\in\mathcal{N}(i)} p^r_{ij} = 1$ and $p^r_{ij} = 0$ whenever i and j are not connected in the PIN. For $(p^r_{ij})_{j\in\mathcal{N}(i)}$ to be a probability distribution we require that $p^r_{ij} \geq 0$, which is guaranteed if $w^r_{ij} \geq 0$. The choice of the transformation of the Pearson correlation w_{ij} must be made carefully to ensure non-negativity and that interpretation of a high edge weight, indicative of an increased likelihood of interaction between connected proteins, is valid.
As explored in Section 2.1, we consider two possible phenotype level stochastic matrices, corresponding to transformations

$$w^r_{ij} = |C^r_{ij}|$$

and

$$w^r_{ij} = (1 + C^r_{ij}).$$

The former definition biases information flow along paths of highly positively or negatively correlated proteins, whereas the latter biases only positively correlated protein paths,

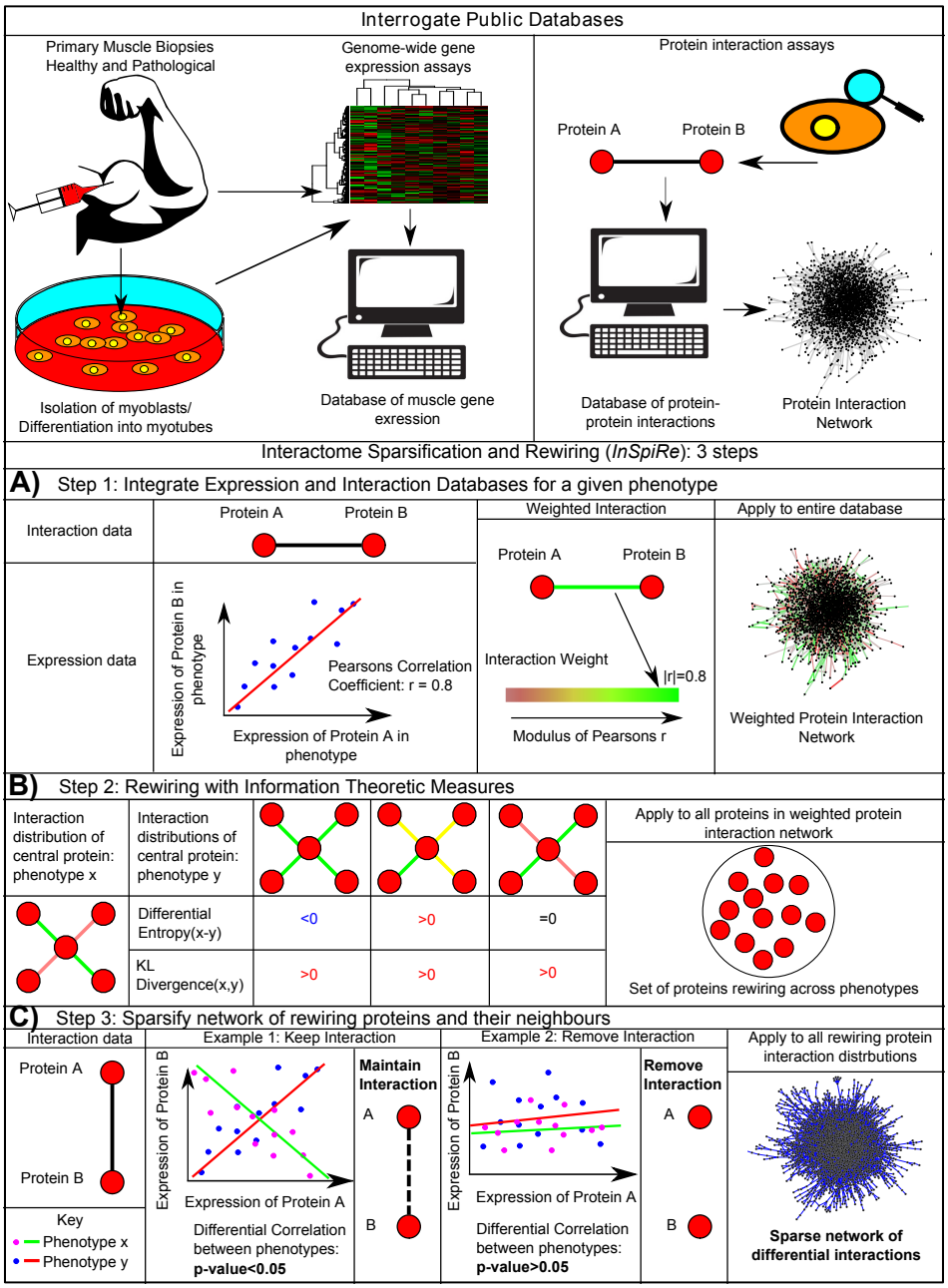

Figure 2.10: **An overview of the InSpiRe algorithm.** Public databases are interrogated for expression and interaction data. (A) Step 1: integration of expression and interaction data, via Pearson correlations, results in a weighted network for each phenotype. (B) Step 2: information theoretic measures detect proteins rewiring between the phenotypes, which are reconnected to their neighbours in the PIN (C) Step 3: sparsification of the rewired network, via differential correlations, results in a sparse network of proteins and interactions which are significantly rewiring.

with negatively correlated being selected against. We explore these two weightings in the context of FSHD and control muscle gene expression in Chapter 5 and find that the weighting based on absolute correlations provides the basis of a more robust discriminator of network rewiring.

2.4.2.2 Step 2: Detecting Rewiring Hotspots - Local Entropy and Kullback-Leibler divergence The second stage of InSpiRe appeals to a local analogue of signalling entropy, namely *local entropy* (defined earlier) to detect rewiring between two phenotypes described by the stochastic matrices constructed by Step 1 of InSpiRe. In addition to local entropy, we also employ a symmetrised Kullback-Leibler divergence to detect a wider class of network rewiring events (Fig 2.10B).

Local entropy quantifies the disorder of a protein interaction distribution in a given phenotype and is based on Shannon entropy.

Given a stochastic matrix corresponding to a phenotype r, p^r_{ij}, the local entropy of protein i is defined as

$$S^r_i := -\frac{1}{\log k_i} \sum_{j \in \mathcal{N}(i)} p^r_{ij} \log p^r_{ij} \tag{74}$$

where k_i is the degree of protein i in the PIN. $S^r_i \in [0,1]$ is a measure of how close a protein's interaction distribution is to uniform. Values close to 0 imply protein i has a deterministic interaction distribution and values close to 1 suggest a uniform profile.

We compute vectors describing the local entropy of each protein in the interactome for two phenotypes. These are statistically analysed using the jackknife technique (below) to identify proteins with significantly different local entropies between phenotypes (significance assessed at the 5% level). Such proteins can be considered as altering heterogeneity in their interaction distributions.

Previous studies into cancer have revealed that proteins with lower local entropy in a given phenotype can be interpreted as being more active in that phenotype [17, 23]. Examples of interaction distribution changes which lead to either an increase or a decrease in local entropy are displayed in Fig 2.10B.

It is possible, however, for a protein interaction distribution to be rewired without changing uniformity and hence without altering local entropy (Fig 2.10B). Consequentially, we introduce another measure based upon Kullback-Leibler divergence, which though lacking the functional interpretation of local entropy is sensitive to such rewiring.

The Kullback-Leibler divergence was introduced earlier and is defined as follows:

Given two probability distributions $P : \mathcal{X} \to [0,1]$ and $Q : \mathcal{X} \to [0,1]$, provided $\{x \in \mathcal{X} : Q(x) > 0\} \subset \{x \in \mathcal{X} : P(x) > 0\}$ the Kullback-Leibler divergence between

P and Q is given by:

$$D_{KL}(P||Q) := \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}. \tag{75}$$

Note that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ as $D_{KL}(P||Q)$ quantifies the expected number of bits required to describe a sample from the probability distribution P given one incorrectly assumes the sample follows the distribution Q. To compare distributions describing the interactions of proteins in control and pathological samples, the choice of the direction of Kullback-Leibler divergence is ambiguous. Arguably both orientations are informative. The quantity $D_{KL}(\text{pathological}||\text{control})$ may describe the amount of information the causal genetic insult of the pathology bestows to the interaction distribution of a protein. Conversely $D_{KL}(\text{control}||\text{pathological})$ may quantify the amount of information required as input to the pathological protein interaction distribution (say by a drug) to restore a healthy phenotype. We will therefore use the symmetrised Kullback-Leibler divergence defined by:

$$D_S(P, Q) := D_{KL}(P||Q) + D_{KL}(Q||P). \tag{76}$$

Given two phenotypes r and s described by stochastic matrices p^r_{ij}, p^s_{ij}, the local symmetrised Kullback-Leibler divergence of protein i between phenotypes is given by

$$K_i(r, s) := D_S((p^r_{ij})_{j \in \mathcal{N}(i)}, (p^s_{ij})_{j \in \mathcal{N}(i)}) \tag{77}$$

$$= \sum_{j \in \mathcal{N}(i)} (p^r_{ij} - p^s_{ij}) \log \frac{p^r_{ij}}{p^s_{ij}}. \tag{78}$$

We note that $K_i(r, s) \in [0, \infty)$ and values near 0 indicate that the interaction distribution of protein i is similar across the phenotypes, large values imply rewiring of protein i between the phenotypes.

We again employ the jackknife technique (below) to determine which proteins have a significantly non-zero Kullback-Leibler divergence (and therefore are rewiring) between two phenotypes (significance assessed at the 5% level).

To create a relevant subset of the PIN proteins identified as significantly rewiring between two phenotypes, are connected to their neighbours in the interactome.

2.4.2.3 Step 3: Sparsification of Relevant Subset of PIN The final step of InSpiRe is the sparsification of the relevant subset of the PIN constructed in Step 2 on InSpiRe, to delete interactions which are not driving the detection of proteins by information theoretic measures (Fig 2.10C). Interactions in this sub-network connecting proteins whose expression is not significantly differently correlated between the two phenotypes (assessed again at the 5% level, via the jackknife) are deleted, sparsifying the network to leave only rewired interactions.

2.4.2.4 Statistical significance determined via the jackknife The jackknife re-sampling procedure has been employed previously to analyse differential local entropies [23], and is considered superior to the bootstrap estimation (for our purposes) which is known to artificially inflate correlations [23, 267].

Jackknife estimation is performed as follows: given a quantity X (e.g. a differential local entropy), we first estimate X from our entire data set, consisting of n samples, denoting this estimate by $\hat{X}$. We next compute n subsequent estimators $(X_i)_{i=1}^n$, from the data set, by removing each sample, one at a time, and re-estimating X. We then compute an estimate for the mean X_μ and the variance X_σ of X via:

$$X_\mu := n\hat{X} - \frac{(n-1)}{n}\sum_{i=1}^{n} X_i \tag{79}$$

$$X_\sigma = \frac{Var[n\hat{X} - (n-1)X_i]}{n-1}. \tag{80}$$

Utilising these estimates we construct a Z statistic

$$Z = \frac{X_\mu}{\sqrt{X_\sigma}} \sim \mathcal{N}(\mu, 1). \tag{81}$$

which can be used to test the hypotheses on the mean of the quantity X. In the InSpiRe algorithm X will either be a differential local entropy a differential correlation or a Kullback-Leibler divergence, hence the null hypothesis will always be that the mean of the quantity if 0. Statistical significance is assessed at the 5% level.

2.4.3 Summary

In this short section we have outlined InSpiRe a local algorithm, which employs an analogue of signalling entropy alongside Kullback-Leibler divergence and statistical sparsification to investigate network rewiring between two phenotypes described by expression data. This algorithm scales the global power of signalling entropy to identify genes and proteins driving changes in this global measure. Moreover, unlike NTE, the InSpiRe algorithm does not require extensive simulation and thus is computationally inexpensive to implement on large networks.

We note that a robust validation of InSpiRe requires a biological context, wherein experimental validation can take place, hence we postpone our validation to Chapter 5, where we explore InSpiRe in detail in the context of FSHD.

2.5 Discussion

The aim of this chapter was to construct and develop entropy based network theoretic tools, which could provide insight into the global and local basis of network rewiring. To

this end we have introduced three such tools and motivated further work.
The first was NTE, a local transfer entropy measure capable of detecting the directed amount of information transferred between any two vertices in a weighted biochemical network. We demonstrated the validity of the NTE measure to detecting differential pathway activation in a phosphorylation network. In addition to being a powerful local measure, we also explored how the theoretical framework behind NTE can be interpreted in the context of a metric space, which can permit the elucidation of global network rewiring properties. In particular, we demonstrated that deformation of network edge weights is followed by a congruent deformation of network dynamics as assessed by a weighted random walk. This global property indicates that a network drug strategy is coherent within the NTE framework.
We next introduced signalling entropy, as a global measure of disorder in protein interaction signalling, computable from a single sample of genome wide gene expression data. We noted that this measure has been previously employed as a discriminator of oncogenic status and here motivated and validated two possible reasons behind this. Firstly, genes which correspond to highly connected vertices in the PIN have been shown to be over-expressed and mutated in cancer [2]. We thus demonstrated theoretically that signalling entropy is more sensitive to the expression of such genes. Subsequently we considered two independent data sets describing gene and protein expression for a number of cancers, empirically confirming that high degree gene expression was a more dominant driver of signalling entropy variation than low degree gene expression. Secondly, tumours are typically more cellularly heterogeneous than healthy tissues. We thus derived a condition for signalling entropy to be elevated on average in samples corresponding to heterogeneous mixtures of cell types. By considering an expression data set describing 33 homogeneous tissue types we confirmed that the condition was valid for all 528 pairwise mixtures, thus providing evidence that signalling entropy is a correlate of intra-sample heterogeneity at the population level. The final network theoretic tool introduced in this chapter was the InSpiRe algorithm, derived from consideration of a local analogue of signalling entropy. This algorithm is intended to extract a subset of the human interactome which is rewiring across two phenotypes, and is more computationally tractable than NTE.
These three network theoretic tools can benefit from further development. In the case of NTE, there is considerable scope for the investigation of high level theoretical properties of network rewiring, such as persistent network states and network dynamic evolution. However, a primary goal of this thesis is the understanding of complex pathology, through the characterisation of cell differentiation and the positing of therapeutics. It is not immediate that further theoretical investigations of the NTE framework will progress these

goals.

The theoretical results regarding signalling entropy, however motivate further investigation to achieve the goals of this thesis. The fact that signalling entropy correlates with intra-sample heterogeneity suggests that the measure may associated with the increased population level diversity of pluripotent stem cells and thus may represent a powerful correlate of cell potency. We will investigate this notion in detail in the next chapter. In addition, intra-tumour heterogeneity is a difficult to measure prognostic factor in epithelial cancer and thus signalling entropy may prove important in survival studies. We will therefore investigate our measure as a prognostic indicator in Chapter 4.

This InSpiRe algorithm is a local analogue of the potentially powerful signalling entropy, designed to provide detailed information about genes and pathways perturbed in pathology, in an easily computable manner. Establishing the validity of this algorithm requires further work and is best performed in a biological context, where it can be compared with other algorithms and where experimental techniques can be employed to validate potentially novel findings. Our algorithm will likely provide greatest insight in a pathology where network analysis is lacking and we thus apply and validate InSpiRe in Chapter 5 in the context of FSHD.

2.6 Materials and Methods, Chapter 2

2.6.1 Estimating NTE

NTE is formulated as the JSD between two probability distributions, for which we can derive closed form expressions. The evaluation of these closed form expressions can be computationally expensive, however, especially if there are multiple vertices of large degree. This is because a main step in the evaluation of the expressions is constructing the set $\mathcal{A}$ of possible signal dynamics over a single time step, which for a network on N vertices is of dimension $\prod_{i=1}^{N} k_i$, where k_i is the degree of vertex i. Moreover, the time complexity of evaluation scales exponentially with the path length parameter n.

For most networks, however, estimation of the probability distributions involved in the NTE expression can be done efficiently. As the model underlying these distributions is a discrete time Markov chain, with a discrete state space, we can employ Monte Carlo simulation for any ISD to provide realisations of the signal distribution on the entire network, for any path length parameter n. From these realisations the probability of a specified signal level at vertex i, given an ISD and path length parameter n, can be estimated as the proportion of simulations in which the specified level is achieved.

Two major considerations need to be addressed to ensure accurate estimation from this procedure. Firstly, it is clear that the more simulations of the model performed, the more accurate the estimate of the probability distribution, moreover the estimate computed from K simulations will converge to the true distribution as $K \to \infty$. Thus it is essential to select K sufficiently large to ensure that the estimated distribution is sufficiently near the true distribution with high probability. Secondly, given a specified K it is important to establish how the error in estimating the probability distributions translates to error in estimating the NTE.

To address the first issue, we consider only the full network *i.e.,* without any vertices set to absorbing state, as the stochastic matrix for the full network will have the fewest deterministic vertices, and thus will be the hardest to estimate probability distributions for. For each probability to be estimated we construct a trace plot describing the change of the estimate with the number of simulations K. This plot allows us to assess convergence of the estimate as K is increased. We select the number of simulations K for each network as the maximal K such that, the shape of the trace plots indicates convergence and the estimates (for every vertex signal probability) at K and $K - 100$ differ by no more than 0.01. The selected value of K is highly dependent on network size, however we found that a value of 500 is sufficient even for very large networks (> 8000 vertices).

To address the second issue of error in the NTE after selecting K, we computed multiple (R) estimations of the signal probability distributions for the full network from K simulations. We then computed, for every $\binom{R}{2}$ estimate pair of the signal probability distribution at a given vertex i, the JSD between the two estimates of the signal distributions. This JSD, computed for vertex i, can be interpreted as the NTE from an arbitrary vertex j to i, when j sends no information to i, if the estimation is perfect, this quantity should be zero. As the estimation is imperfect we obtain $\binom{R}{2}$ estimates of the error in the NTE, deriving from error in the estimation of the probability distributions from simulation, for each receiving vertex. From these estimates we can approximate the first two moments of the error distribution, a NTE for a given receiving vertex is considered not attributable to error, provided it lies at least two standard deviations above the maximal error observed for the vertex.

A program for implementation of the above described algorithm for estimating NTE was written in R and is included as supplementary material to our publication [268].

2.6.2 Assigning an ISD to the biological network

To compute the NTEs over the two perturbations of the biological signalling network we must first define an ISD and edge weights for each perturbation, from the data. As the

kinases in the network must be phosphorylated to phosphorylate their direct targets, two connected proteins with highly positively correlated phosphorylation levels across single cell observations under a given perturbation, are likely interacting. Thus a suitable edge weight which captures the strength of a phosphorylation interaction represented by an edge (i, j) under a given perturbation k is $1 + C_{ij}^k$, where C_{ij}^k is the Pearson correlation of phosphorylated protein levels across single cell measurements under perturbation k.
Defining the ISD is less trivial and requires consideration of the question we wish to answer and some technicalities. To determine the differences in information transfer between the two perturbations, it makes sense to consider an ISD which quantifies the difference in phosphorylated protein levels between the two treatments. A technical issue to consider is how different signal distributions on the same network lead to different NTE values between vertices. It is clear that weighting a vertex with a non-zero signal leads to a higher NTE value between that vertex and its downstream interaction partners (depending on the value of n) than weighting the vertex with zero signal. Thus vertices with an information deficit in one perturbation versus another should be weighted with zero signal in that perturbation and a non-zero signal in the other.
A more subtle issue concerns the number of unique signal values attainable at each vertex for a given ISD and how this relates to the NTE. It is somewhat intuitive that if the inputs to a given vertex each have a unique initial signal value, then the range of values attainable by the receiving vertex will be more diverse than if all the inputs had the same initial signal value. Thus one may hypothesise that the NTE from one vertex to another vertex, with multiple inputs, will be larger if the input vertices have unique initial signal values than if they have identical signal values.
To explain this concept, consider the network shown in Fig 2.11, consisting of one central vertex with 3 possible inputs, which are each as likely to forward signal to the central vertex as they are to forward signal to a separate independent neighbour. We consider the effect of the ISD on the NTE from the circled vertex to the central vertex for path length parameter $n = 1$ via two cases. In the first scenario the initial signal at every input vertex is identical (Fig 2.11A, in this example this signal value is one), while the signal at every other vertex starts at zero. In the second case every input vertex is given a unique value (Fig 2.11B, in this example these values are 1, 2 and 3), while the other vertices are again given initial signals of 0. Consider the probability distribution of signal at the central vertex after a single signal transfer event. For the first ISD, this distribution can take 4 unique values (namely 0, 1, 2 and 3), however for the second ISD, with unique signal values at the inputs, the distribution can take 7 values (integers 0 through 6). Now consider the signal distribution at the central vertex given that we prevent the circled

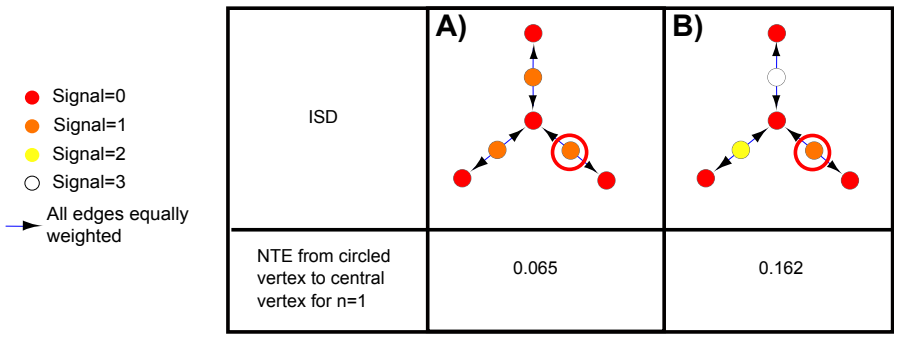

Figure 2.11: **NTE dependence on the ISD.** Comparing the NTE from source vertices to a target vertex when the ISD at source vertices are (A) non-unique and (B) unique, note that a unique ISD at input vertices results in a higher NTE from source to target.

vertex from sending information. For the first ISD, this distribution now admits only 3 possible values (0, 1 and 2), while for the second ISD only 4 possible values are now attainable (0, 2, 3 and 5). Thus the size of the "coding alphabet" at the central vertex after removal of input from the circled vertex has shrunk from 4 to 3 under the first ISD but from 7 to 4 under the second ISD (a much greater fall). Consequentially, we notice that the NTE from the circled vertex to the central vertex is lower for the first ISD than the second ISD.

If we follow this concept that inputs with unique initial signal send more information to their outputs, it is logical that vertices with significantly higher signal in one perturbation versus another should be given a unique initial signal value in the first perturbation, reflecting their capacity to send more information about the network.

All that remains now is to consider how to assign initial signal to vertices which do not display a great difference in signal distribution across the two perturbations of our biological network. One solution to this problem is to assign all these vertices an identical initial signal, in this way they can transfer more information than vertices with a signal deficit in one perturbation versus another but less information than vertices with a signal surplus.

Guided by these concepts we constructed the ISD for each perturbation of the phosphorylation network as follows. We utilised the limma package in R [269] to compute t-values testing, for each vertex in the network, the hypothesis that the phosphorylated protein level of the vertex was significantly different in the two treatments. If for a given vertex the phosphorylated protein levels were significantly lower ($p < 0.05$) in one perturbation versus another it was assigned an initial value of zero in that perturbation and a unique

initial value (here chosen as the absolute t-statistic of the test) in the other perturbation. All vertices which did not display significant changes between the two perturbations were assigned the same non-unique initial value of 1 in both perturbations.
The ISDs and edge weights for the two perturbations of the biological network are provided in Fig 2.3A-B.

2.6.3 mRNA Expression data

The analysis in this chapter required the use of two microarray gene expression data sets. For the analysis of signalling entropy's association with high degree vertices, we considered data from 1980 primary breast cancer patients by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) project [157]. The data consists of microarrays profiled on the Illumina HT 12v3 platform and log normalised as described in [157]. This expression data is accompanied by extensive clinical annotation describing tumour grade, size, stage, cellularity, histological subtype, lymph node status, hormone receptor status, Pfam50 subtype and p53 mutation status, as well as patient age at diagnosis, menopausal status, treatment and survival data. As described in [157] the expression data was divided into a discovery set of 997 samples and a validation set of 983 samples, each representing a distribution of clinical variables and molecular subtypes observed in the full data set.
For the analysis of signalling entropy's association with intra-sample heterogeneity we downloaded from the GEO database [270], normalised data corresponding to accession number GSE2361 [265]. This data set profiled 33 distinct healthy human tissues and 3 foetal tissues on the Affymetrix Human Genome U133A Array platform. Data was inter array normalised.

2.6.4 Protein Expression data

Data describing protein expression via immunohistochemistry for 15805 proteins across multiple samples describing 20 cancerous tissues was downloaded from the Human Protein Atlas data base [263]. The expression of each protein in each cancer sample was scored on an integer scale from 1-4, where 1 indicated no expression, 2 low expression, 3 moderate expression and 4 high expression. For each protein we computed the average expression score across samples corresponding to a given cancerous tissue, resulting in a protein expression vector for each of the 20 cancerous tissues.

2.6.5 Protein Interaction Network

The PIN utilised in the signalling entropy analysis was downloaded Pathway Commons (`http://www.pathwaycommons.org`) [8] (date stamp 13th June 2012). Using this comprehensive resource, we built an integrated network including the Human Protein Reference Database [271], the National Cancer Institute Nature Pathway Interaction Database (NCI-PID) (`pid.nci.nih.gov`), the Interactome (Intact) (`http:/www.ebi.ac.uk/intact/`) and the Molecular Interaction Database (MINT) (`http://mint.bio.uniroma2.it/mint/`). Protein interactions in this network include physical stable interactions such as those defining protein complexes, as well as transient interactions such as post-translational modifications and enzymatic reactions found in signal transduction pathways, including 20 highly curated immune and cancer signaling pathways from NetPath (`http://www.netpath.org`) [272]. Redundant interactions were removed and only genes with an EntrezGene ID annotation were retained.

Interactions derived from publicly available interaction databases have been criticised, for false positives introduced by the lack of cellular organisation constraints in Y2H assays. To remove network edges which may likely represent false positives, a sparsification procedure based on imposing a (bi-directional) signalling hierarchy on the network. This sparsification was kindly performed by Andrew E. Teschendorff (supervisor for this project). Briefly, the procedure removed edges between proteins whose main cellular localisations are not adjacent in the context of the signalling hierarchy by utilising GO annotations [26]. Genes may be annotated to multiple domains, and thus mutually exclusive sets were constructed by favouring the more external domain, so that a gene annotated to both extracellular and transmembrane domains was allocated to the extracellular domain only, and a gene annotated to both transmembrane and intracellular domains was assigned to the transmembrane domain only. Thus, all EntrezGene identifiers were annotated to one of the extracellular (EC), transmembrane receptor (MR) or intracellular (IC) domains. Nodes in the PIN which could not be assigned to one of these domains were removed and the resulting network was pruned by removing edges (interactions) inconsistent with the signalling hierarchy structure. Thus, only edges with corresponding end nodes in the following combinations were allowed: EC-EC, EC-MR (or MR-EC), MR-IC (or IC-MR) and IC-IC. This resulted in a maximally connected PIN of 8,434 nodes and over 300,000 interactions.

2.6.6 Integration of the PIN with gene expression data

To integrate expression data with the PIN, expression profiles of probesets/proteins mapping to the same EntrezGene identifier in the PIN were averaged. Proteins in the PIN without probesets/proteins representing their coding genes on the array, were removed. The resulting maximally connected component of the PIN defined an interaction network for each data set.

2.6.7 Signalling Entropy

Signalling entropy for a given sample in a data set is computed as described above in Section 2.3, where it is theoretically analysed. We note, however, that as the PIN for each data set can vary due to the genes described by each experimental technology, it is important to consider the normalisation of signalling entropy, so that it is comparable across data sets.

Briefly, for the computation of signalling entropy each sample is first integrated with the PIN to create a sample specific stochastic matrix, $P = (p_{ij})$, via the mass action principle as described above (Section 2.1).

For each protein i we then define the local entropy of its interaction distribution, S_i, which quantifies the promiscuity of its signalling within the sample:

$$S_i = -\sum_{j \in \mathcal{N}(i)} p_{ij} \log p_{ij}. \tag{82}$$

Signalling entropy is then computed from the entire stochastic matrix p_{ij} as the entropy rate, $\tilde{SR}$, of the stochastic process described by p_{ij}:

$$\tilde{SR} = \sum_i \mu_i S_i, \tag{83}$$

where μ_i denotes the stationary distribution of the stochastic matrix, satisfying $\sum_i \mu_i p_{ij} = \mu_j$. We note that μ_i is therefore the non-degenerate eigenvector of P corresponding to the eigenvalue 1 and that by the Perron Frobenius theorem, the existence of μ_i requires that the matrix P be irreducible; this is guaranteed by the fact that the PIN considered is connected and non-bipartite [273].

The maximum entropy rate of a weighted network, M_R, depends solely upon its adjacency matrix, $A = (A_{ij})$, and can be calculated as the entropy rate of the stochastic matrix $p_{ij} = A_{ij}\nu_j/\lambda\nu_i$, where λ and ν are the dominant eigenvalue and corresponding eigenvector of A, respectively [274]. In order to ensure the signalling entropy values are comparable across data sets, we present our data driven findings in terms of normalised

signalling entropy:

$$SR = \tilde{SR}/M_R. \tag{84}$$

R-scripts for the computation of signalling entropy are freely available for download at `www.sourceforge.net/projects/signalentropy`.

3 Signalling Entropy as the energy potential of Waddington's Landscape

3.1 Introduction

In our introductory chapter we motivated the need for network theoretic tools which can elucidate global principles of biological network rewiring. Furthermore, we posited that there likely exists an organisation to such network rewiring during complex biological processes, which manifests via the predictable change of a global measure of biological signalling as the process evolves. The development of such a measure, however, requires the selection of a biological process to investigate and we subsequently motivated cellular differentiation as a clear candidate.

We explained how organisms develop from the differentiation of a single pluripotent ES cell, whilst adult tissues are maintained by a tissue specific stem cell pool, ready to expand and differentiate to restore what is lost. We also revealed the incredible sophistication of the orchestrated gene expression regimes required for cellular differentiation, making it hardly surprising that aberrant differentiation underpins the most complex and malicious pathologies. We examined two antithetical such pathologies in detail, namely breast cancer and FSHD. We saw how considerable evidence has arisen suggesting that tumours may possess a small population of CSCs from which the tumour bulk differentiates [167], and how such cells are posited to be responsible for the growth and repair of the tumour in the same way that healthy stem cells contribute to the growth and maintenance of the organism. Further we saw how in FSHD mis-expression of a transcription factor, *DUX4*, is believed to produce a subtle perturbation to the muscle differentiation programme resulting in atrophic or disorganised muscle fibres in patients.

In our examination of the mathematics of cellular differentiation we saw how recently, it has been proposed that pluripotency, is an emergent statistical property of cell populations [54, 98, 80], not well-defined at the single-cell level. Specifically, it has been argued that high cellular heterogeneity underpins the pluripotent or multipotent capacity of stem cell populations, with differentiated cell populations representing a more homogeneous synchronised state [54].

In Chapter 2, we developed a number of network theoretic tools. In particular, we examined signalling entropy, the entropy rate of a single sample, mass-action principle derived stochastic matrix, describing a random walk model of network traffic. Signalling entropy is a sample specific measure of disorder in interactome signalling, however, we further revealed, by theoretical investigation that our measure also correlates with intra-sample heterogeneity at the population level.
Motivated by this, we here investigate signalling entropy in the context of cellular differentiation. We first explain the rationale of using signalling entropy in this context, before demonstrating by the consideration of genome wide gene expression data that the measure is elevated in pluripotent and multipotent cells as opposed to differentiated tissue, across multiple cell fates. We subsequently consider time course gene expression profiling of differentiation across multiple lineages, demonstrating that signalling entropy systematically decreases throughout such time courses. We next consider CSCs and demonstrate that signalling entropy is elevated in these cells as compared to the tumour bulk. Finally we consider local entropy and demonstrate this measure capable of detecting the key drivers of the differentiation process. We also find that signalling entropy is a more robust indicator of differentiation potential than other gene expression based indicators. These results mark out signalling entropy as a powerful measure of the differentiation potential of a single sample and motivates the use of our measure in a pathological context. We note in particular that signalling entropy can be considered as a proxy for the energy potential in Waddington's differentiation landscape.

3.2 Results

3.2.1 Rationale of signalling entropy as a measure of differentiation potential

The rationale behind signalling entropy as a correlate of differentiation potential is based on the proposition that our measure captures the average level of signalling pathway promiscuity in a given sample, in a manner related to intra-cellular pathway activation diversity and inter-cellular heterogeneity. Undifferentiated and plastic cells, such as stem cells, which must maintain diverse pathway activation and heterogeneous populations, would thus be characterised by a state of high signalling entropy. Similarly, since differentiation results in the activation of a restricted set of molecular signalling pathways and a homogenisation of cell populations, one would expect a reduction in signalling entropy as this process progresses. As explored in Chapter 1, a commonly used model for cell differentiation potential is Waddington's landscape, in which stem cells are considered to occupy states of high potential energy, which decreases throughout cellular differentiation.

If our model is correct then signalling entropy represents a proxy for the energy potential in Waddington's landscape (Fig 3.1A).

We have demonstrated analytically in Chapter 2 that signalling entropy is an average measure of inter-cellular heterogeneity. By its nature as an entropy rate, signalling entropy can also be expected to measure the level of intra-cellular pathway diversity, in homogeneous cell populations. To confirm this we devised a simulation model in which the signalling entropy of an unbiased random walk on our PIN (*i.e.* $p_{ij} = 1/k_i$ where k_i is the degree of node i) representing a promiscuous, poised, intra-cellular signalling state, is compared to the signalling entropy obtained by randomly activating individual genes and specific pathways (Fig 3.1B, Materials and Methods, Chapter 3). In the case where individual genes were activated, this led, in approximately 70% of perturbations, to a reduction in signalling entropy (Binomial $p < 0.001$, Fig 3.1B). However, in the case where whole signalling pathways were activated, a reduction signalling entropy was observed in 85% of cases (Binomial $p < 10^{-10}$, Fig 3.1B). These results confirm that signalling entropy is a measure of intra-cellular pathway diversity, which decreases on the deterministic activation of genes and pathways.

Thus by approximating intra-cellular and inter-cellular heterogeneity in pathway activation, signalling entropy can be considered an appropriate quantifier of the changes in signal transduction that occur during cellular differentiation.

3.2.2 Signalling entropy is raised in pluripotent stem cells

Given our theoretical results we next sought to determine if our measure could discriminate pluripotent samples from more differentiated ones in a data driven manner. We evaluated the signalling entropy of all 219 samples in the *stem cell matrix* (SCM), 59 of which were deemed pluripotent, and the remaining 160 non-pluripotent [79] (Materials and Methods, Chapter 3). We observed that signalling entropy was significantly higher in the pluripotent cell lines ($p < 10^{-10}$, Fig 3.2A).

To provide an independent benchmark for signalling entropy we considered a transcriptomic measure of pluripotency, derived by Mikkelsen *et al.* [275] (Materials and Methods, Chapter 3). This pluripotency score was also found to be significantly higher in the pluripotent cell lines, and was significantly correlated, with signalling entropy across samples in the SCM, confirming that our measure associates with other correlates of stemness (Fig 3.2B).

To further validate signalling entropy we considered an independent data set profiling 107 human ES cell and 52 iPSC lines, as well as 32 differentiated tissue samples (SCM2) [276]. Our measure achieved 100% accuracy in discriminating pluripotent from differenti-

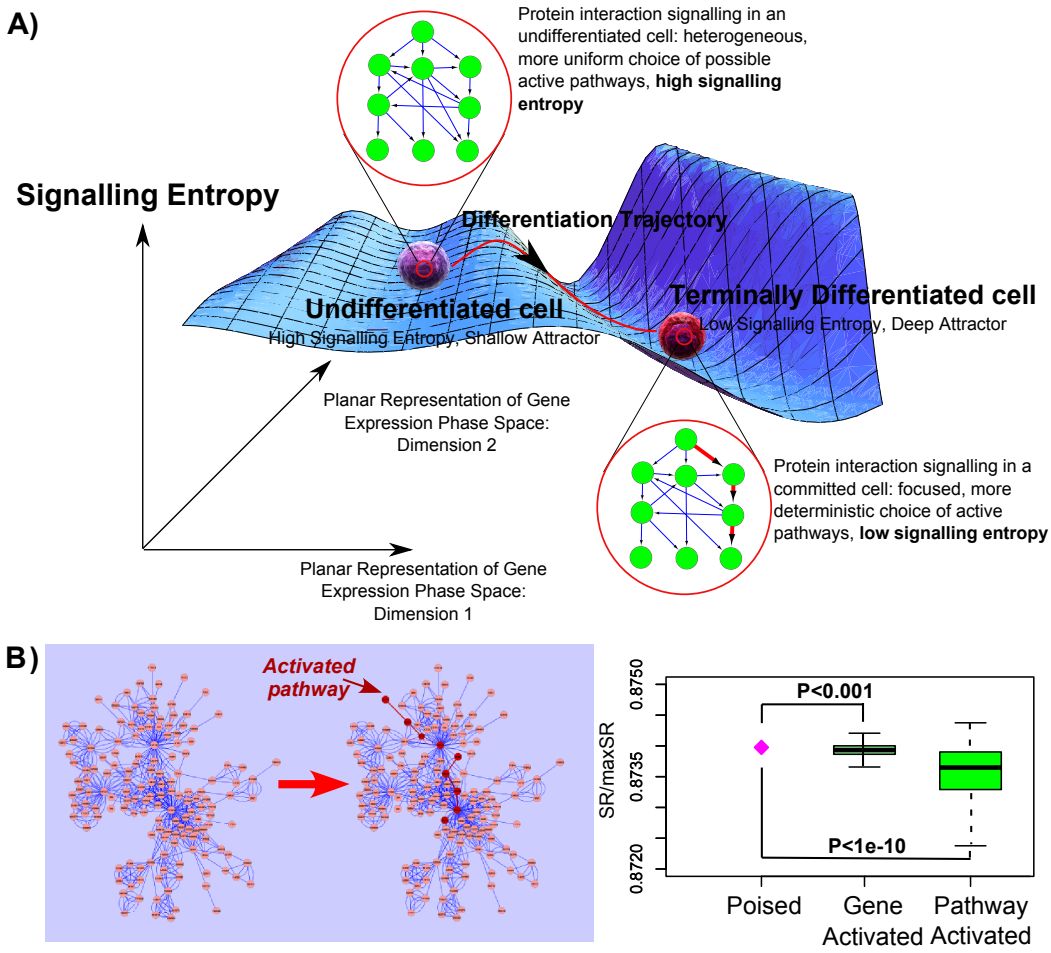

Figure 3.1: **Signalling entropy as the height in Waddington's Landscape.** (A) The z-axis represents the signalling entropy of a cellular sample. The plane spanned by the x-and-y axes represents gene expression phase space. In accordance with Waddington's landscape, an undifferentiated cell occupies a region of phase space with increased potential energy, signalling entropy is considered to be a proxy for this potential. (B) Simulation of pathway activation in a realistic PIN decreases signalling entropy. The boxplot compares the signalling entropy (SR/maxSR) of the unbiased random walk state (diamond), to the signalling entropies obtained by separately activating each individual gene in the network (> 1000 perturbations), and those obtained by activating whole signal transduction pathways (100 pathways). Binomial test p-values are given.

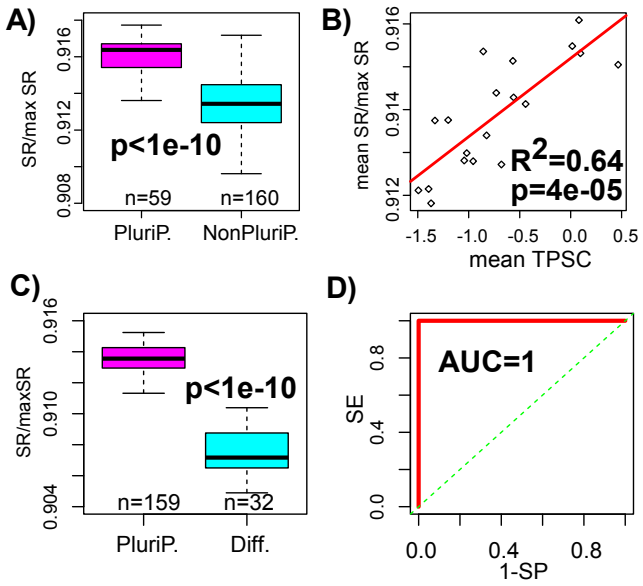

Figure 3.2: **Signalling entropy correlates with the differentiation potential of a sample.** (A) Signalling entropy (SR/maxSR, y-axis) is elevated in the 59 pluripotent compared to the 160 non-pluripotent cell-lines from the SCM compendium, the p-value is from a Wilcoxon rank sum test. (B) Signalling entropy correlates with the pluripotency score of Mikkelsen *et al.*, (values for replicates of each cell type have been averaged). Linear regression p-value and R^2 value are given. (C) Signalling entropy (SR/maxSR, y-axis) is elevated in the 159 pluripotent samples compared to the 32 differentiated samples from the SCM2, the p-value is from a Wilcoxon rank sum test. (D) Corresponding ROC curve plus AUC of signalling entropy discriminating pluripotent from differentiated cells in SCM2.

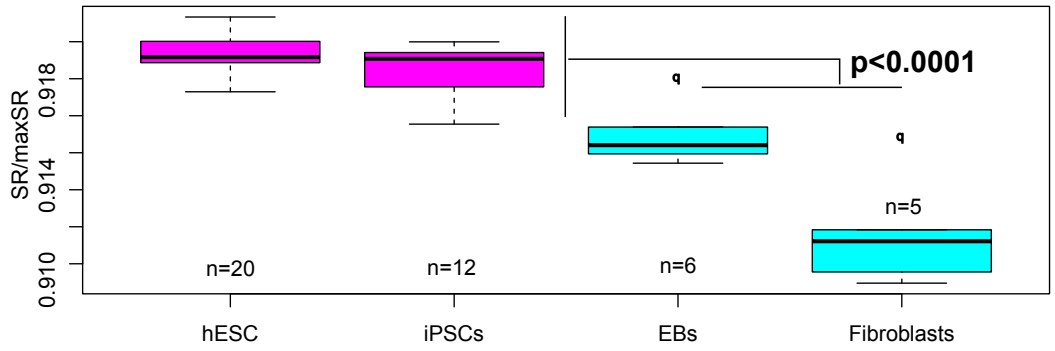

Figure 3.3: **Signalling entropy is similar in ES cells and iPSCs.** Signalling entropy displays similar values for ES cells and iPSCs, significantly higher than the values corresponding to embroid bodies (EB) and iPSC parental fibroblasts. This is in line with the similar differentiation potential of these pluripotent cell types.

ated samples in this data set (Fig 3.2C-D). We also observed that iPSC samples exhibited high signalling entropy values, similar to those of ES cells, and significantly higher than those of their parental differentiated cells ($p < 0.0001$, Fig 3.3).
We note that cell cycle gene expression has previously been associated with ES cells [277]. To determine if the expression of these genes was driving signalling entropy's discriminatory power, we removed cell proliferation and cycling genes from our PIN [277] and recomputed signalling entropy over the remaining maximally connected component. We found that our measure remained a significant discriminator of pluripotent cells and hence is not determined purely by cell proliferation (Fig 3.4).

3.2.3 Signalling entropy is raised in adult stem cells

Following the finding that signalling entropy is raised in pluripotent stem cells, we next considered our measure evaluated over multipotent adult stem cell types, including neural stem cells (NSCs), hematopoietic stem cells (HSCs) and mesenchymal stem cells (MSCs). We found that all these adult stem cell types exhibited higher signalling entropy values than their differentiated progeny but lower values than pluripotent ES cells (Fig 3.5). Thus, signalling entropy is capable of discriminating cells within a lineage according to their differentiation hierarchy.
To test this finding further in the context of a single lineage, we considered a haematological data set [278], encompassing a number of different blood cells, including differentiated (*e.g.* monocytes), and less differentiated types (*e.g.* *CD34*+ HSCs and erythrob-

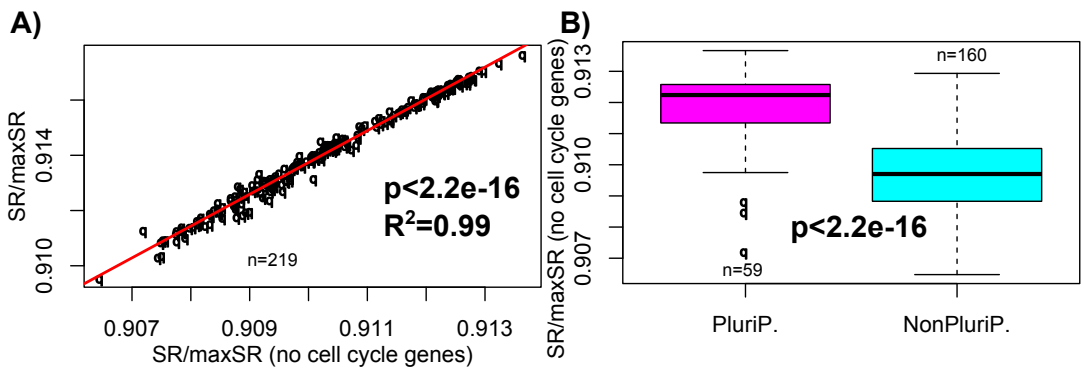

Figure 3.4: **Signalling entropy is not driven by cell cycle genes.** (A) Signalling entropy shows near identical variations between samples in the SCM after removal of cell cycle genes. Linear regression R^2 and p-values are provided. (B) Signalling entropy computed without cell cycle genes is still a powerful discriminator of pluripotency in the SCM, the p-value corresponds to a Wilcoxon rank sum test.

lasts/megakaryocytes). We found that signalling entropy recapitulated the hematopoietic differentiation hierarchy consistent with prior knowledge [279, 280] (Fig 3.6).

3.2.4 Signalling entropy dynamically decreases during differentiation time courses

We have demonstrated that signalling entropy is elevated in pluripotent cells as compared to multipotent cells and in multipotent cells as compared to differentiated cells, across multiple cell fates. However, these cell types represent discrete positions in a lineage and we wonder whether signalling entropy dynamically decreases during cell fate transition.
To investigate this we first considered time course expression data of differentiated retinal pigment epithelium (RPE) cells, which were induced to de-differentiate, for expansion, and then were re-differentiated back in to RPE (Materials and Methods, Chapter 3). Elegantly, signalling entropy increased dynamically upon de-differentiation, reaching a maximum at the point the cells were induced to re-differentiate, after which it systematically decreased as the cells reverted to RPE (Fig 3.7A).
As another example, we considered a time course data set consisting of human promyelocytic leukaemia progenitor (HL60) cells, differentiating into neutrophils [83] via two independent stimuli (ATRA and DMSO). In both cases, signalling entropy was significantly reduced over the differentiation time course (ATRA stimulus, $R^2 = 0.96$, $p < 10^{-8}$, Fig 3.7B).

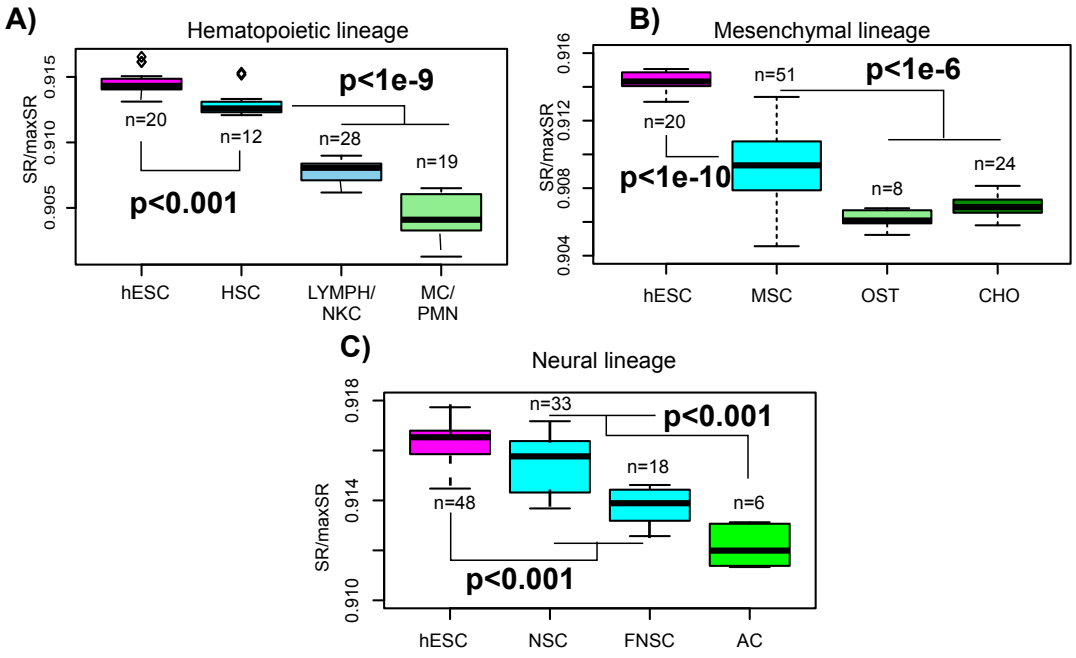

Figure 3.5: **Signalling entropy associates with the differentiation hierarchy within multiple lineages.** (A) Signalling entropy associates with the differentiation hierarchy in the hematopoetic lineage, being highest in ES cells, then adult HSCs and lowest in differentiated T & B-cell lymphocytes plus natural killer cells (LYMPH/NKC), and monocytes plus neutrophils (MC/PMN). (B) Signalling entropy associates with the differentiation hierarchy in the mesenchymal lineage being highest in ES cells, then adult MSCs and lowest in differentiated osteoblasts (OST) and chondrocytes (CHO). (C) Signalling entropy associates with the differentiation hierarchy in the neural lineage being highest in ES cells, then adult NSCs and foetal neural stem cells (FNSC) and lowest in differentiated primary astrocytes (AC). Wilcoxon rank sum test p-values.

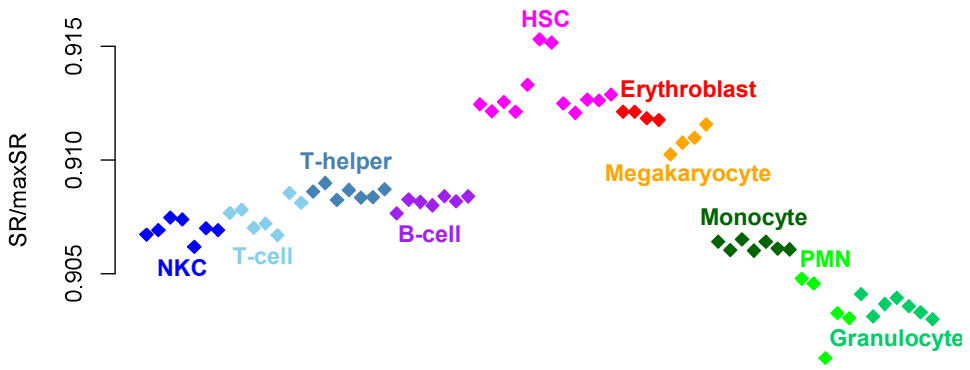

Figure 3.6: **Comparison of signalling entropy across major blood cell types in the hematopoietic system.** Blood cell types have been arranged with lymphoid cells to the left, and myeloid (monocytes & granulocytes) cells to the right, with the pluripotent HSCs and less differentiated erythroblasts/megakaryocytes in the middle. Signalling entropy is lower for the more differentiated lymphoid and myeloid lineages. The HSCs and PMNs (polymorpho neutrophils) were profiled with the Affymetrix U133 Plus 2 platform, while the rest of samples were profiled with Illumina Human WG6v2 arrays. The similarity of signalling entropy for PMNs and ganulocytes confirms the cross platform validity of our measure.

As yet further confirmation we considered a data set describing ES cell differentiation into pancreatic β-cells at 5 time points, in sextuplicate [70]. Once again, we found that signalling entropy systematically decreased as the cells differentiated (mean of sextuplicates: $R^2 = 0.68, p = 0.04$ Fig 3.7C).
Thus signalling entropy dynamically decreases as cells differentiate into diverse cell fates. It is worth noting that these dynamic changes were independent of cell-proliferation (*i.e.* the results held when cell-cycle genes were removed from the network).

3.2.5 Signalling entropy discriminates cancer stem cells from the tumour bulk

As discussed in Chapter 1, cancer is a disorder of development and there are several hypotheses surrounding oncogenesis. One hypothesis is that cancerous tissue arises via de-differentiation of healthy tissue into a more proliferative, heterogeneous state. In line with this theory, it has previously been demonstrated that signalling entropy, computed over transcriptomic data, is elevated in cancerous tissue compared to its normal counterpart [262]. Here we further validate this finding using proteomic data from the Human Protein Atlas describing 20 different malignancies and healthy tissues [263]. We find that signalling entropy is still elevated at the protein level in cancerous as opposed to healthy tissue ($p < 0.003$ Fig 3.8A).
Recently, considerable evidence has arisen suggesting that CSCs may drive malignancy, given signalling entropy's power as a discriminator of stem cells in a healthy context, we analysed an expression data set profiling putative CSCs and their parental tumours, across a number of different tissues [281]. This revealed that CSCs exhibited a marginally higher signalling entropy than their non-stem like counterparts (Fig 3.8B).

3.2.6 Dynamic changes in local entropy identifies key differentiation genes and pathways

Given the power of signalling entropy as a measure of a sample's differentiation potential we next investigated whether we could identify critical mediators of the differentiation process through the consideration of local entropy. As an entropy rate, signalling entropy is defined as a weighted sum of local entropies, corresponding to given proteins in the PIN. A low local entropy of a protein indicates a deterministic interaction distribution, which can be attributed to involvement of the protein in an active pathway. We thus hypothesised that decreases in the local entropy of certain genes during differentiation may indicate their activation during this process.

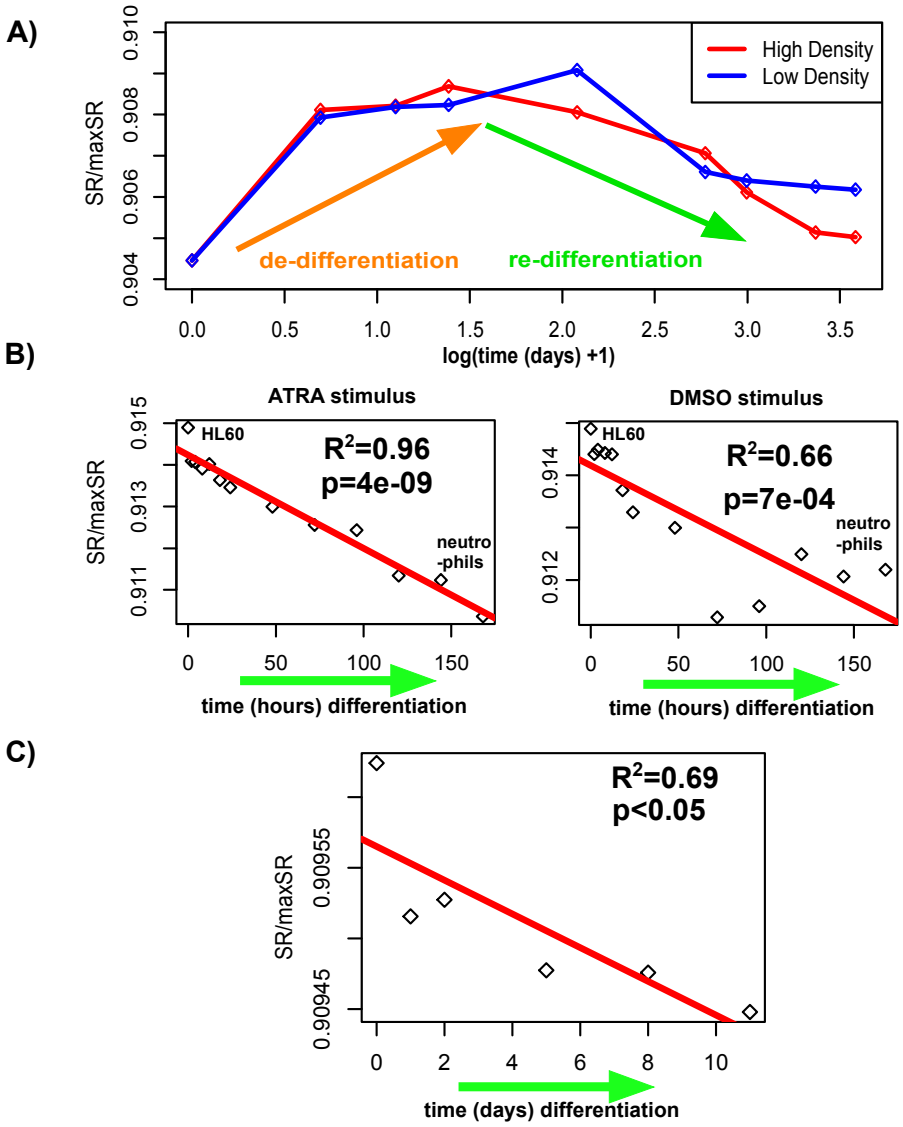

Figure 3.7: **Signalling entropy exhibits dynamic changes during the differentiation process.** (A) Signalling entropy (SR/maxSR, y-axis) first increases then decreases in a time course de-differentiation followed by re-differentiation of RPE plated at two different densities. (B) Signalling entropy systematically decreases during HL60 leukemic progenitor cell differentiation into neutrophils, stimulated by two independent initial stimuli: ATRA and DMSO. (C) Signalling entropy systematically decreases during ES cell differentiation into pancreatic β-cells (points displayed are the average of sextuplicates). For (B) and (C) we provide the R^2 values and associated p-values for a linear regression of signalling entropy against differentiation time.

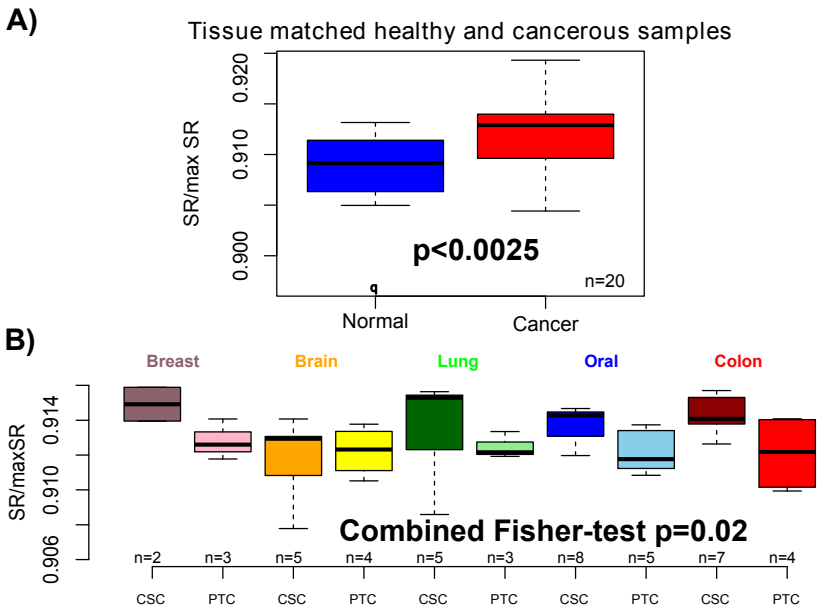

Figure 3.8: **Signalling entropy is higher in cancerous tissue and in CSCs.** (A) Signalling entropy computed over proteomic data describing 20 cancerous and matched healthy tissues is significantly higher in cancerous tissue, p-value is from a paired Wilcoxon test. (B) Comparison of signalling entropy between putative CSCs and their parental tumour cell lines (PTC) for five different tissue types shows signalling entropy is higher in CSCs. The combined Fisher t-test p-value is given.

We considered, as a proof of principle, the case of *Notch*-signalling. *Notch* signalling is inactive yet inducible in the pluripotent state, with activation normally associated with differentiation [282, 283, 284, 285, 286, 287]. We thus anticipated that essential components of the *Notch* signalling pathway would exhibit a higher local entropy in the pluripotent, compared to the differentiated state. Considering data from the SCM [79], we were able to confirm a decreased local entropy for 10 *Notch* pathway genes in our PIN, in non-pluripotent samples (Figs 3.9 & 3.10). To confirm the statistical significance of this, in none of 10000 random selections of 10 genes from the PIN did we observe the same level of consistency and statistical significance as for the *Notch* pathway genes ($p < 0.0001$). This result indicates that reduced local entropy of the *Notch* pathway is a key feature of the non-pluripotent state, confirming that our measure can detect pathway activation during differentiation.

To further test the added value of local entropy, we revisited the HL60 differentiation time course data [83]. Using linear regression we identified the genes showing the most significant decreases in local entropy during the differentiation process. Ranking genes according to those showing the largest reductions in signalling entropy and performing a subsequent Gene Set Enrichment Analysis (GSEA), we identified *JAK-STAT* signalling as one of the key pathways (Fig 3.11 & 3.12). The involvement of this pathway in neutrophil differentiation is heavily supported by previous studies [288, 289, 290, 291].
To confirm the statistical significance of the *JAK-STAT* pathway finding and the biological relevance of the structure of the PIN, we computed local entropies after randomly permuting the gene expression profiles over the vertices of our network. This led to no significantly enriched biological terms (adjusted p-values > 0.05), emphasising the criticality of the network structure to our results. Finally, we note that other non-network based approaches fail to identify the *JAK-STAT* pathway (Fig 3.11).

3.3 Discussion

The aim of this chapter was to propose and test the hypothesis that signalling entropy, a network theoretic measure of intra-cellular signalling promiscuity and inter-cellular heterogeneity, is correlated with the cell potency of a sample.
We computed the signalling entropy of over 1000 samples, describing gene expression in healthy cell types from many diverse lineages and differentiation stages, as well cancerous tissues and CSCs, profiled using a variety of different technologies. Through this analysis we have demonstrated that signalling entropy provides a near absolute quantification of the differentiation potential of any given sample.
In the context of normal physiology, ES cells and other pluripotent cell types were cor-

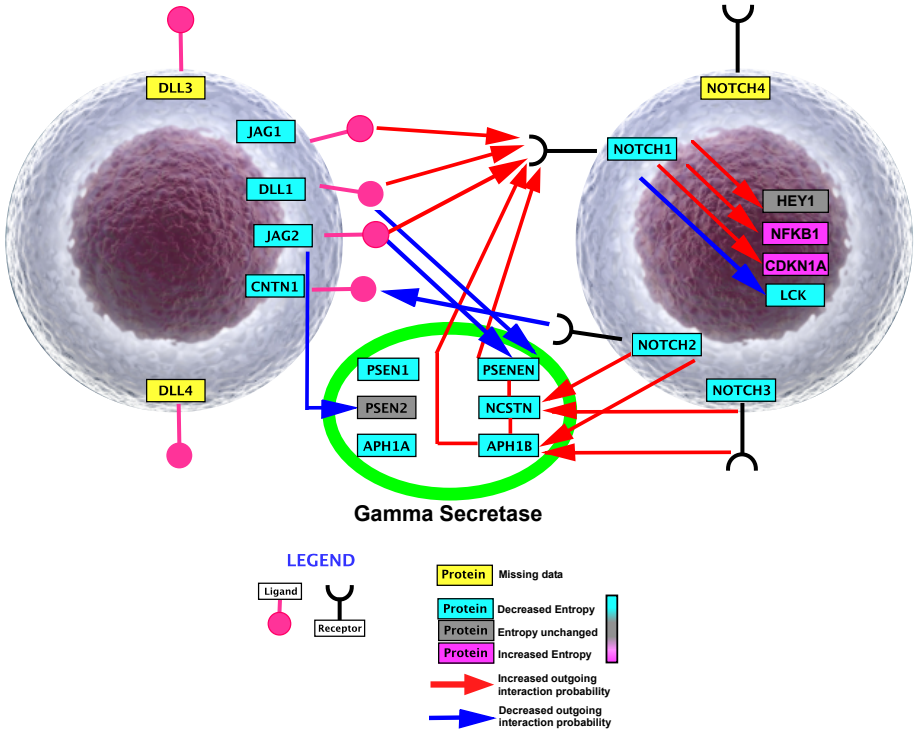

Figure 3.9: **Local entropy identifies rewiring of *Notch* pathway mediators in non-plutipotent cells.** Graphical rewiring diagram of alterations in the interaction preferences in the main *Notch* pathway components after transition from pluripotency. The main increased and decreased probability interactions are depicted as arrows. Most of the *Notch* pathway components have reduced local entropies, highlighting the fact that the interaction distributions become more focused (mostly increased interactions between *Notch* pathway members).

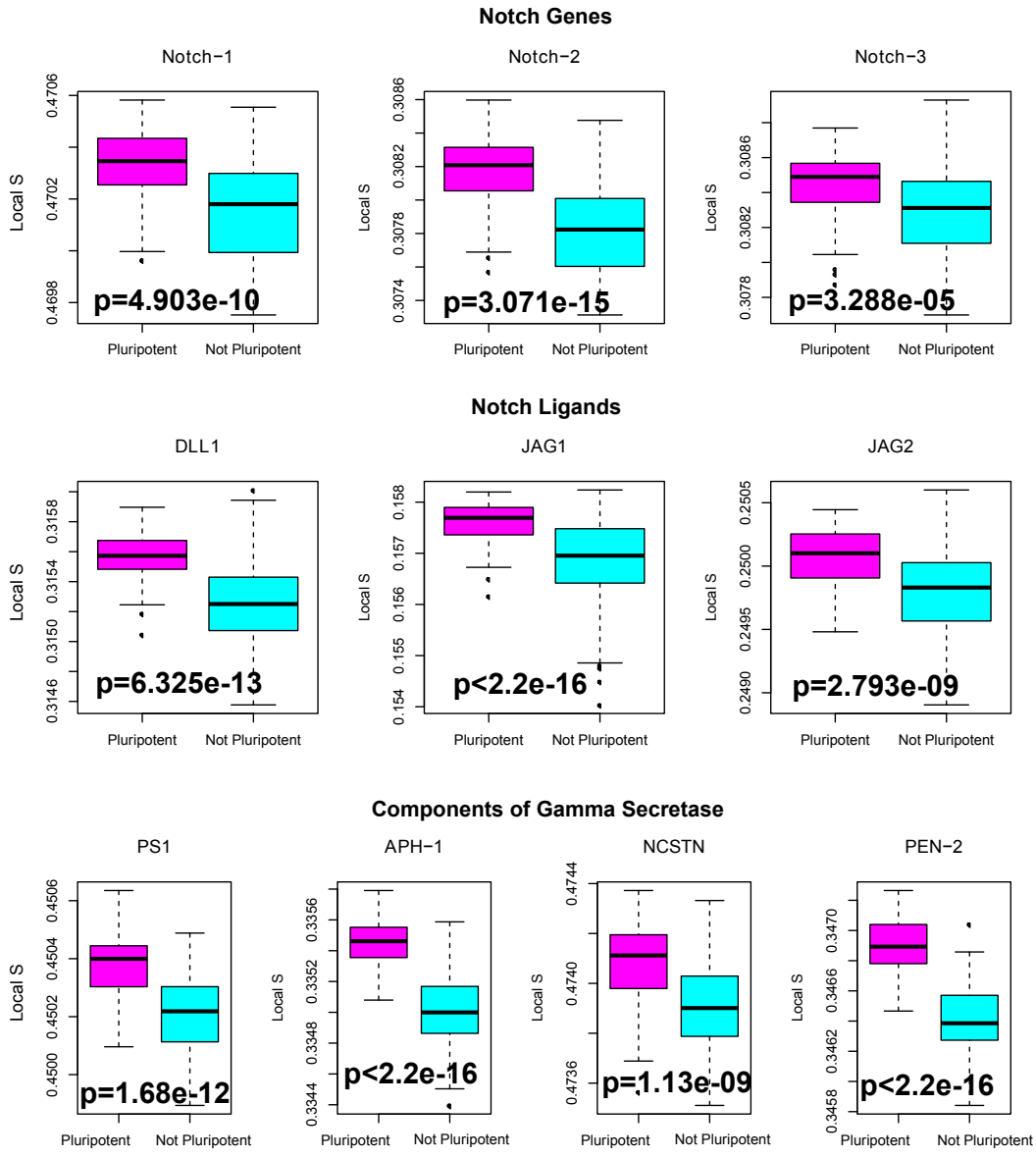

Figure 3.10: **Local entropy decreases in critical *Notch* pathway mediators in non-plutipotent cells.** Boxplots showing the significant reduction in local entropy in 10 critical *Notch* pathway components after transition from pluripotency. p-values are from a Wilcoxon test.

Negatively Correlated Local Entropy Methodology			
GO ID	GO Term	P Value	Adjusted P value
GO:0042517	positive regulation of tyrosine phosphorylation of Stat3 protein	6.72E-07	0.0010
GO:0050731	positive regulation of peptidyl-tyrosine phosphorylation	1.65E-06	0.0013
GO:0004896	cytokine receptor activity	4.70E-06	0.0013
GO:0042516	regulation of tyrosine phosphorylation of Stat3 protein	5.20E-06	0.0026
GO:0042531	positive regulation of tyrosine phosphorylation of STAT protein	6.68E-06	0.0025
GO:0019838	growth factor binding	1.16E-05	0.0016
GO:0046427	positive regulation of JAK-STAT cascade	1.28E-05	0.0039
GO:0042509	regulation of tyrosine phosphorylation of STAT protein	2.97E-05	0.0075
GO:0002673	regulation of acute inflammatory response	3.10E-05	0.0067
Mar and Quackenbush Regression Methodology; Core gene set			
GO ID	GO Term	P Value	Adjusted P value
GO:0003676	nucleic acid binding	1.08E-06	0.00833
GO:0016070	RNA metabolic process	5.52E-06	0.0129
GO:0044446	intracellular organelle part	7.10E-06	0.0129
GO:0044422	organelle part	7.24E-06	0.0129
GO:0044428	Nuclear part	8.41E-06	0.0129
GO:0006350	transcription	3.18E-06	0.0407
GO:0032774	RNAb iosynthetic process	5.46E-06	0.046
GO:0032991	macromolecular complex	5.60E-06	0.046
GO:0006694	steroid biosynthetic process	5.74E-06	0.046
Positively Correlated Expression Methodology			
GO ID	GO Term	P Value	Adjusted P value
GO:0005773	vacuole	1.67E-06	2.30E-04
GO:0005764	lysosome	9.66E-07	2.66E-04
GO:0000323	lytic vacuole	9.66E-07	2.66E-04
GO:0044437	vacuolar part	5.78E-04	0.052
GO:0006952	defense response	5.59E-05	0.089
GO:0006955	immune response	2.47E-04	0.19
GO:0005886	plasma membrane	0.0035	0.21
GO:0043020	NADPH oxidase complex	0.011	0.38
GO:0015629	actin cytoskeleton	0.0089	0.39
GO:0005774	vacuolar membrane	0.013	0.41

Figure 3.11: ***JAK-STAT* signalling is significantly enriched among genes which decrease their local entropy during HL60 differentiation into neutrophils.** GSEA results for genes which showed a significant reduction in local entropy over the two stimulations of HL60 differentiation into neutrophils, reveals an enrichment of *JAK-STAT* signalling. In contrast the non-network based regression analysis of Mar and Quackenbush, and the consideration of gene expression correlation with differentiation, identifies only high level processes.

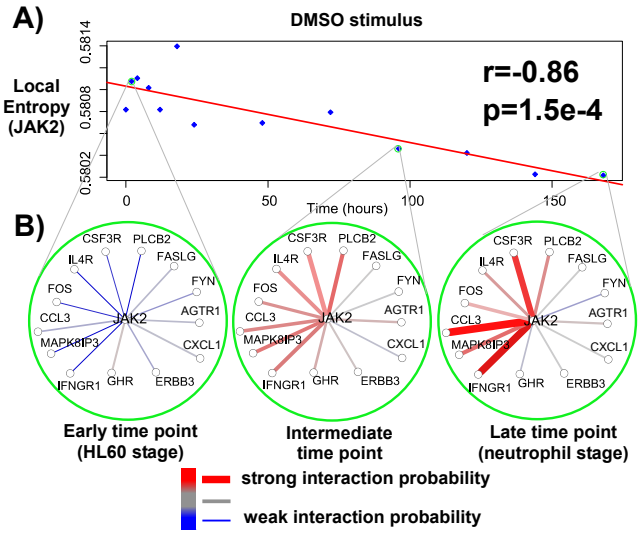

Figure 3.12: **Local entropy detects the rewiring of *JAK2*'s interaction distribution into an active state during neutrophil differentiation.** (A) The local entropy of *JAK2* systematically decreases during neutrophil differentiation (here shown is DMSO stimulated differentiation, the result is similar for ATRA stimulated). Pearson's r and corresponding p-value for linear regression are displayed. (B) We see that the reduction in local entropy is due to the interaction profile of *JAK2* rewiring into a more active state during differentiation.

rectly predicted to exhibit the highest signalling entropy, followed by multipotent stem cells (*e.g.* NSC/HSC/MSC), with terminally differentiated cells exhibiting significantly lower signalling entropy. In the context of cancer, CSCs exhibited higher signalling entropy than the tumour bulk, although this difference appears substantially less than between normal stem cells and their differentiated progeny. We also demonstrated that the local entropy of proteins whose activation is critical to differentiation decreased as the process progressed.

As demonstrated in Chapter 2, signalling entropy is a measure of intra-sample heterogeneity. Our results thus support the view that the pluripotent state is a statistical property of a diverse cell population [97, 54]. It will be important to assess, when single cell data becomes available, whether the increased signalling entropy in pluripotent cells requires a population level assessment, or whether the pluripotent state is achieved via an inherent intra-cellular stochasticity. [97, 54].

Our measure thus provides a very general system's level correlate of the undifferentiated state of a sample. This is in contrast to pluripotency gene expression signatures [275, 292], which lack a systems level interpretation and understanding. We also note that such signatures could only consistently discriminate pluripotent from non-pluripotent cell types, but generally failed to discriminate cell types located further down the differentiation hierarchy ([293] analysis performed by project supervisor: Andrew E. Teschendorff, data not shown). Thus, signalling entropy provides a more refined classification of distinct cell types across the global differentiation hierarchy than previous transcriptomic measures.

Moreover, though we observed some variation in signalling entropy between studies profiling the same cell types, we note that these variations were in general very small. Thus signalling entropy provided a relatively robust measure of cell potency across studies and technologies, with ES cells always demonstrating the highest levels. This robustness stems from two key features. Firstly, signalling entropy is self-calibrating, as it is constructed from ratios of gene expression intensity values, making it a dimensionless quantity, fairly insensitive to experimental technology or normalisation method. This is in contrast to pluripotency gene expression signatures, which showed significant variations between studies. Secondly, unlike pluripotency expression signatures signalling entropy does not depend on feature selection, thus it is not subject to over-fitting and independent of tunable parameters.

In summary, we have here demonstrated that signalling entropy, correlates with the differentiation potential of a cellular sample and can be considered a proxy for the energy potential in Waddington's landscape. To this end we have achieved one of the main goals of the thesis, namely the development of an entropic network theoretic tool, which can

approximate the differentiation level of a single sample and hence form the basis of a methodology designed to investigate developmental pathology. In the next two chapters we will thus investigate signalling entropy in the context of cancer prognostics and FSHD pathomechanisms.

3.4 Materials and Methods, Chapter 3

3.4.1 Signalling Entropy

Signalling entropy was computed as described in the Materials and Methods, Chapter 2.

3.4.2 Simulation analysis of signalling entropy as a measure of pathway promiscuity

Our simulation study was intended to investigate the effect of perturbations to the stochastic matrix p_{ij} on signalling entropy, to ascertain whether gene and pathway activation can lower our measure as postulated.

Our perturbations alter the stochastic matrix in such a way as to prevent the validity of the closed form expression derived in Chapter 2. Consequentially, signalling entropy for a perturbed stochastic matrix must be computed by numerical approximation of the left eigenvector of the stochastic matrix, a computationally expensive procedure. Without loss of generality, we therefore randomly sampled 2000 genes from the PIN, resulting in a maximally connected component consisting of ~ 1500 genes. This reduced network size allowed for swift numerical computation of signalling entropy under a large number of perturbations.

To simulate a pluripotent poised state, in which all pathways can be activated with equal likelihood, a stochastic matrix was defined as an unbiased random walk: $p_{ij} = 1/k_i$, whenever $j \in \mathcal{N}(i)$, 0 otherwise. We considered two types of perturbation to this state, the activation of genes and the activation of pathways. In the first case, we considered all vertices of degree ≥ 2 in the PIN, for each such vertex, a randomly picked edge was assigned a large weight (in the range $0.8 - 0.95$), with the remaining edges assigned a low weight (~ 0.1), ensuring that the sum of weights equals 1, as required for a stochastic matrix. This was done for each vertex in the network separately, resulting in > 1000 perturbations and associated signalling entropies. In the second scenario, pathway activation, we randomly selected sequences of connected vertices (paths) and assigned edges connecting vertices in a path a high weight as above. Path lengths were chosen to be of maximum length 9, corresponding to the diameter of the network, and cycles were not permitted. We simulated 100 distinct active pathways and computed the resulting

signalling entropies.

3.4.3 Gene expression data

We compiled a transcriptomic expression database, consisting of $\sim$ 1000 samples, from the sources described below. The choice of data sets was guided by our desire to perform specific comparisons between certain cell types, thus requiring that these specific cell types were profiled as part of the same study, in order to minimise potential confounding by batch effects. In the case of Affymetrix data, if RMA normalised data was publicly available, this was used. If normalised data was not provided, or if this was unspecified, raw data was downloaded and RMA normalised using the Bioconductor affy package. In the case of Illumina data sets, the normalised data provided by the respective studies was used. In all cases, quantile normalisation was performed to normalise across all arrays within a study. Finally, for later integration with the PIN, expression profiles of probes mapping to the same EntrezGene identifier were averaged. Probes mapping to multiple genes were excluded.

3.4.3.1 The stem cell matrix (SCM and SCM2) compendia We obtained the normalised gene expression data of the SCM compendium, consisting of 219 samples (59 pluripotent, 160 non-pluripotent) [79]. The pluripotent samples consisted of 48 ES cells, 5 teratocarcinomas, 3 iPSCs and 3 germ tumour cell samples. All the SCM data were generated using Illumina Human Reference8 arrays.
In addition, we also obtained the normalised gene expression data (Illumina HT12v3 expression arrays, GSE30652) describing a subset of the samples used in the SCM2 compendium, consisting of 107 ES cell lines, 52 iPSC samples and 32 samples from differentiated tissue [276].

3.4.3.2 Further ES cell, MSC and iPSC data sets We obtained normalised data from GEO [270] for two studies profiling human ES cell lines and derived MSCs. One study [294] profiled an ES cell line (H1) at 3 different passages, giving 3 replicates, as well as two MSC precursors and a bone marrow derived MSC population. These 6 samples were profiled on Affymetrix HG-U133A arrays. The other study [295] profiled two ES cell lines and derived MSCs, all 4 samples were performed in triplicate resulting in 12 samples, profiled Affymetrix HG-U133 Plus 2 arrays.
We also downloaded the normalised data for three additional, independent studies profiling bone-marrow derived MSCs (MSC-BM), all profiled on Affymetrix HG-U133 Plus 2

arrays [296, 297, 298]. Samples however varied in terms of donor, donor age and passage numbers. GSE7888 [296] consisted of a total of 23 MSC-BM samples, GSE9593 [297] of 13 MSC-BM samples, and GSE9520 [298] of 6 MSC-BM samples.
In addition, we downloaded the normalised data for three further independent, studies profiling ES cell samples [299, 300, 301], on the Affymetrix HG-U133 Plus 2 array: GSE7896 (5 samples) [299], GSE13828 (5 samples) [300], and GSE15148 (10 samples) [301]. GSE13828 also profiled iPSCs from skin fibroblast samples taken from a child with spinal muscular atrophy [300]. Specifically, in addition to the 5 ES cell samples, there were 3 iPSC samples and 2 skin fibroblast samples. GSE15148 contained 16 iPSCs derived using episomal vectors from 2 foreskin samples [301]. Another fourth data set comparing iPSCs ($n = 12$) to parental fibroblasts ($n = 6$), including ES cells ($n = 20$) was obtained from [302]. The raw data, generated with Affymetrix HG HT U133A arrays, was quality checked and RMA normalised. Thus, all these data sets allowed a three-way comparison between adult differentiated cells, the iPSCs derived from them, and ES cells.

3.4.3.3 Expression data describing the differentiation of MSCs into osteoblasts and chondrocytes We downloaded the normalised data for two studies [303, 304] profiling MSC samples and differentiated osteoblasts and chondrocytes, profiled on Affymetrix HG-U133 Plus 2 arrays.

3.4.3.4 Combined haematological data set We obtained the normalised data from GEO for three haematological studies. One study profiled five *CD34* + HSC samples and five differentiated neutrophil samples using the Affymetrix Human Genome U133A Plus 2 array [305]. The remaining two studies profiled 3 and 4 *CD34* + HSC cell samples respectively using the same Affymetrix platform [306, 307]. We also obtained normalised data from the HaemAtlas [278]. The array for this study was the Illumina Human WG-6 v2 Expression beadchip. However, both array platforms led to integrated networks of similar size encompassing effectively the same genes, permitting direct comparison.

3.4.3.5 Time course de-differentiation and re-differentiation experiment of RPE cells The RPE de-differentiation and re-differentiation normalised data set was obtained from ArrayExpress (E-MTAB-854), where a detailed experimental protocol can be found. Briefly, a single cell suspension of RPE was derived from ES cells and plated in 96 well plates at two densities in triplicate, a high density (100,000 cells/cm^2) and a low density (8000 cells/cm^2), and cultured for 5 weeks. RNA was extracted from the starting RPE cell suspension and from the plated cells at 8 subsequent time points (1,

2, 3, 7, 15, 19, 29 and 35 days after plating). In the high density plates, cells proliferate and de-differentiate (acquiring a mesenchymal morphology) during the first 5 days before re-differentiating to RPE by the end of the 5 weeks. In the low density plates cells proliferate and de-differentiate for the first 7 days before re-differentiating to RPE by end of the 5 weeks. RNA was profiled on Illumina HumanHT12v4 arrays.

3.4.3.6 Time course HL60 neutrophil expression data We obtained from GEO, the raw CEL files for the microarray (Affymetrix Human Genome U95 Version 2 Array) dataset collected by Huang *et al.* [83] describing HL60 progenitor cells differentiating into neutrophils. The dataset consists of 25 samples; the first sample is of HL60 progenitor cells during proliferation and the remaining 24 samples correspond to two time courses, each of 12 time points (2 hrs, 4 hrs, 8 hrs, 12 hrs, 18 hrs, 1 day, 2 days, 3 days, 4 days, 5 days, 6 days, 7 days), describing differentiation of HL60 progenitors into neutrophils. Each time course corresponds to differentiation initiated via stimulation with a given medium; either DMSO or ATRA supplementation. The dataset underwent quality control assessment via the arrayQualityMetrics package in R, and was normalised using RMA.

3.4.3.7 Time course pancreatic β-cell differentiation The pancreatic β-cell differentiation data set was obtained from ArrayExpress (E-MTAB-817), where a detailed experimental protocol can be found. Briefly ES cells were plated in sextuplicate, (duplicates of 3 thawed batches) and induced to differentiate into pancreatic β-cells, RNA was isolated at 5 time points (0, 1, 2, 5, 8 and 11 days) and profilled on Illumina HumanHT-12 v3.0 Expression BeadChip arrays.

3.4.3.8 Cancer stem cell and parental cancer cell data set In order to compare putative CSC to their parental tumour cell PTC lines we used the normalised data (GEO) from [281]. Specifically, we focused on five tissues: breast with 2 CSCs and 3 PTCs, brain with 5 CSCs and 4 PTCs, lung with 5 CSCs and 3 PTCs, oral cavity with 8 CSCs and 5 PTCs and colon with 7 CSCs and 4 PTCs. In the case of colon we used positivity of CD133, a colon stem cell marker to assign CSC/PTC status (samples on GEO are likely to have been mis-labelled). Reported results across all tissue types is however independent of inclusion or exclusion of the colon subset. All samples were profiled on Affymetrix HG-U133 Plus2 arrays.

3.4.4 Protein Expression data

Data describing protein expression via immunohistochemistry for 15805 proteins across multiple samples describing 20 cancerous and matched healthy tissues was downloaded from the Human Protein Atlas data base [263]. The expression of each protein in each sample was scored on an integer scale from 1-4, where 1 indicated no expression, 2 low expression, 3 moderate expression and 4 high expression. For each protein we computed the average expression score across samples corresponding to a given tissue, resulting in a protein expression vector for each of the 40 tissues.

3.4.5 Gene expression based pluripotency signatures

We considered two pluripotency gene expression signatures: A 19 gene pluripotency signature, consisting of 11 up and 8 down regulated genes in pluripotent cells described by Mikkelsen *et al.*, [275] and a 189 gene pluripotency expression signature derived by Palmer *et al.* [292].
For the Mikkelsen *et al.* signature, the pluripotency score of a given sample was derived as the t-statistic comparing the expression levels of the 11 up-regulated genes to the 9 down-regulated ones.
For the Palmer *et al.* signature a score was derived from the 189 genes following the *principal component analysis* (PCA) procedure outlined in [292]. We performed a focused PCA on the SCM compendium, utilising the genes from the Palmer signature represented in the array platform as a feature space (a total of 157 genes). The top principal component from the PCA correlated with pluripotency status (Wilcoxon rank sum, $p < 10^{-10}$), validating the signature in the SCM data. The sign of the loadings in the top principal component was then utilised to assign the 157 genes into up and down-regulated categories. A pluripotency score of an independent sample was then obtained as the t-statistic of the test comparing the expression levels of the up and down regulated genes.

3.4.6 Protein Interaction Network

The PIN utilised in this Chapter was constructed as described in the Materials and Methods, Chapter 2.

3.4.7 Integration of PIN with gene expression data

To integrate the gene expression data with the PIN, expression profiles of probesets and proteins mapping to the same EntrezGene identifier in the PIN were averaged. Proteins in the PIN without probesets representing their coding genes on the array, were removed.

The resulting maximally connected component of the PIN defined an interaction network for each array platform. In the case of Affymetrix U133 Plus 2 arrays (the most widely used array in this work), the resulting maximally connected integrated PIN consisted of 8290 nodes and 299459 edges. For the Affymetrix Human Genome U95 v2 arrays, the integrated PIN consisted of 5555 nodes and 175640 interactions. For the Affymetric U133A arrays, the PIN was of size 7027 nodes and 246005 edges. Finally, for the Illumina Human WG-6 v2 expression beadchip, the integrated maximally connected PIN consisted of 8135 nodes and 290360 edges, *i.e.*, very similar in size to the Affymetrix U133 Plus 2 arrays.

3.4.8 GSEA

Functional annotation and GSEA was performed using the DAVID Bioinformatics Resources 6.7 [308]. The proteins in the PIN were used as a background gene set and p-value estimated using Fisher's exact test. Adjusted p-values used the Benjamini-Hochberg correction as implemented in DAVID.

4 Signalling Entropy Correlates with Clinical Outcome in Epithelial Cancer

4.1 Introduction

In the introductory chapter we explored the molecular diversity of breast cancer in detail. We revealed the considerable heterogeneity of this epithelial cancer and how it leads to a multitude of subtypes, each with differing prognosis and treatment options. We also saw how this diversity has led to the rise of gene expression based prognostic assays, which may be employed to stratify patients into different clinical classes. Furthermore, we examined the clonal evolution and CSC hypotheses of tumour development, explaining how acquired mutations and epigenetic alterations can lead to heterogeneity within individual tumours. We motivated the development of network theoretic measures aimed at the understanding of CSCs and how they drive intra-tumour heterogeneity, positing that entropic approaches may prove fruitful.

In Chapter 2 we demonstrated that signalling entropy is a correlate of intra-sample heterogeneity, and hence may prove useful in understanding intra-tumour heterogeneity. Moreover, in Chapter 3 we demonstrated that signalling entropy is a powerful measure of the differentiation potential of a single sample in a healthy context, and is capable of discriminating CSCs from the tumour bulk. Here we investigate signalling entropy in

more detail in the context of a prognostic measure of intra-tumour heterogeneity.
The observation that epithelial cancers display intra-tumour heterogeneity in cell morphology and that this may impact on clinical outcome was made over 40 years ago [309, 310, 266]. It was only recently, however, that intra-tumour heterogeneity was demonstrated at the molecular level [311, 312, 313, 314]. Not only has this revealed considerable diversity in the mutational landscapes of single tumours [312, 314], but also that gene expression based prognostic scores differ in outcome on multiple biopsies from the same tumour [311], raising significant challenges for their clinical application. Other recent studies have shown that cellular heterogeneity may promote progression to neoplasia and dictate response to treatment [311, 315, 316].
The clinical assessment of intra-tumour heterogeneity is currently an open problem, with a number of proposed solutions [313]. Arguably the major clinical concern from intra-tumour heterogeneity is its contribution to therapeutic resistance [317], thus the most relevant measures must consider this carefully. Treatment resistance of heterogeneous tumours likely stems from the ability of the cancer cell population to activate a diverse array of pathways permitting subversion of the intended therapeutic response [318, 318, 314, 317]. Thus it is reasonable that a clinically relevant measure of intra-tumour heterogeneity will quantify the signalling promiscuity of a tumour and infer its phenotypic plasticity, making signalling entropy a strong candidate.
Intra-tumour heterogeneity will have several sources, however as mentioned earlier an important contribution is likely to come from epigenetic diversity, driven by CSCs [314, 161, 162]. Putative breast and lung CSCs have been identified by surface marker expression, isolated, and demonstrated to be chemotherapeutic resistant [167, 165, 155, 319, 320]. However, the measurement of the stemness of a tumour in a manner related to CSCs remains a significant challenge. Although ES cell gene expression signatures have been constructed and shown to be prognostic in breast cancer [277, 64, 321], their overall prognostic significance seems limited and they are unable to distinguish CSCs from the tumour bulk [277]. Given that signalling entropy is a measure of stemness in a healthy context, and is elevated in CSCs, it is not unlikely that our measure may also prove a robust correlate of tumour stemness in a manner related to CSC abundance and hence potential intra-tumour heterogeneity.
Motivated by these results of previous chapters, we here propose and investigate signalling entropy as a prognostic measure of tumour stemness and intra-tumour heterogeneity.
We demonstrate that our measure strongly associates with both histological and transcriptomic measures of tumour anaplasia across a wide variety of malignancies (breast, lung, prostate, glioma), validating our measure as a correlate of tumour stemness.

We next investigate signalling entropy as a prognostic variable in the context of the two leading causes of cancer death world wide: breast cancer and *non-small cell lung cancer* (NSCLC). We find that signalling entropy associates with different clinical subtypes in both malignancies, providing novel hypotheses on pathogenesis.
In NSCLC, we discover that signalling entropy is elevated in patients with strong smoking history. We also find that signalling entropy is higher and less variable in lung *squamous cell carcinoma* (SCC) as compared to lung adenocarcinoma. Our results suggest this difference may be driven by the smoking of high tar, unfiltered cigarettes which is associated with SCC [322].
In breast cancer we find that the luminal B subtype displays a a high signalling entropy, despite a low tumour stemness (judged by the expression of ES cell genes). This suggests that increased intra-tumour heterogeneity may be driving high signalling entropy in luminal B breast cancer.
By performing a meta-analysis over 2000 primary lung adenocarcinoma and 3500 primary breast cancer samples we further demonstrate that signalling entropy is significantly prognostic in both cancers, independently of other clinical variables. Importantly, our measure is prognostic within the *stage I* stratum of lung adenocarcinoma and within ER positive and ER negative breast cancer subsets, in contrast to other transcriptomic prognostic indicators.
Finally, we consider proteomic data from the Human Protein Atlas describing healthy and cancerous tissue from numerous tissue types. We demonstrate that the signalling entropy increase in cancerous tissue correlates with tissue specific cancer mortality rates. Thus we here reveal that signalling entropy is a powerful prognostic factor in epithelial cancer capable of providing novel insights into pathomechanisms.

4.2 Results

4.2.1 Rationale of signalling entropy as a prognostic measure

As shown in Chapter 3, a cellular sample with a high signalling entropy will possess a robust intra-cellular signalling regime, capable of activating a number of diverse pathways. Such a sample will likely be resistant to perturbation by pharmacological agents. Hence a tumour with a high signalling entropy may be expected to possess an innate resistance to therapeutic intervention and thus a poor prognosis. We also demonstrated that signalling entropy is elevated in treatment resistant CSCs as compared to the tumour bulk (Chapter 3), suggesting that a high signalling entropy tumour will likely possess a greater proportion of such cells. As CSCs are considered capable of driving tumour

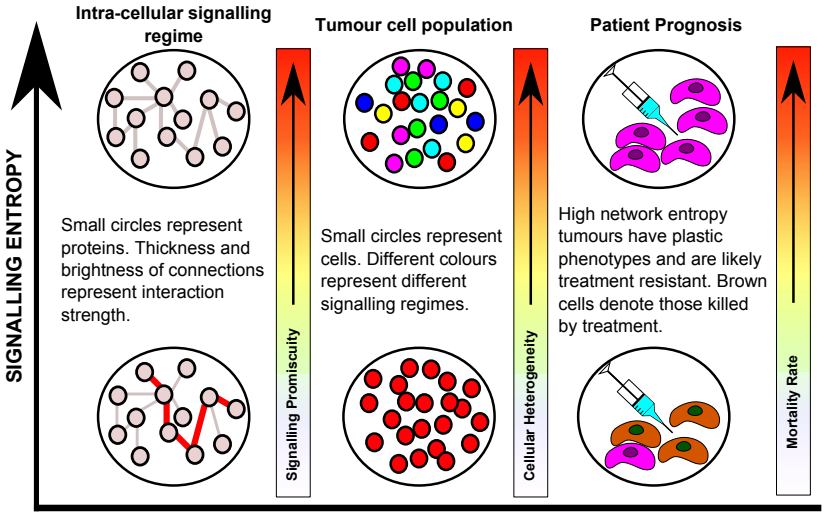

Figure 4.1: **Rationale behind signalling entropy as a prognostic factor in cancer.** A high signalling entropy of a tumour sample indicates a promiscuous, intra-cellular signalling regime and a heterogeneous cancer cell population. The consequence of a high signalling entropy is thus a tumour with a plastic phenotype, capable of activating diverse pathways in response to treatment. High signalling entropy tumours are thus likely to result in higher patient mortality.

growth and heterogeneity, one can therefore expect a high signalling entropy tumour to be of poor prognosis. In addition we saw in Chapter 2, that signalling entropy is elevated in heterogeneous as opposed to homogeneous samples, whence a high signalling entropy tumour can be expected to possess a diverse population of cell types, some of which may be resistant to therapeutic response.

Thus, in contrast to other prognostic measures, signalling entropy can be expected to associate with tumour stemness in a manner related to CSC abundance, intra-cellular signalling promiscuity and intra-tumour heterogeneity, making it good candidate for an improved prognostic indicator (Fig 4.1).

4.2.2 Signalling entropy correlates with tumour stemness

We demonstrated in Chapter 3 that signalling entropy correlated with differentiation potential in healthy tissue. We therefore hypothesised that our measure would correlate

with transcriptomic and histological measures of tumour differentiation across multiple cancer types. To investigate this we considered 4 malignancies: breast cancer, NSCLC, prostate cancer and glioma. We found a strong correlation between signalling entropy and conventional measures of tumour stemness.

4.2.2.1 Signalling entropy correlates with measures of tumour stemness in breast cancer We first considered the METABRIC breast cancer data sets. For this malignancy two appropriate transcriptomic measures of tumour differentiation are the Ben-Porath *et al.* 100 gene ES cell signature [277, 323] and the 97 gene Sotiriou *et al.* grade signature [324].

As expected, we found that both signatures were strongly correlated with signalling entropy in the METABRIC data sets (Ben-Porath: $p < 2.2 \times 10^{-16}$, Sortiriou: $p < 2.2 \times 10^{-16}$, (Fig 4.2A & B) indicating that signalling entropy is indeed associated with external, transcriptomic measures of the stemness of a tumour.

As anticipated, signalling entropy also strongly correlated with histological tumour grade in breast cancer, being significantly higher in grade 3 tumours compared with grade 2 and significantly higher in grade 2 tumours as compared with grade 1 ($p < 7.3 \times 10^{-15}$ and $p < 2.6 \times 10^{-4}$ respectively, Fig 4.2C). Importantly, we found that unlike signalling entropy the Ben-Porath *et al.* signature was unable to distinguish between grade 1 and grade 2 tumours in the discovery set of METABRIC (Ben-Porath *et al.* signature: $p = 0.4$, signalling entropy: $p < 2.6 \times 10^{-4}$, Fig 4.2D). This result indicates that our measure is more sensitive to tumour stemness than ES cell transcriptomic signatures.

We further note that tumours with a high signalling entropy display a bi-modality of enrichment for the Ben-Portah *et al.* ES cell signature. High ES cell enriched, high signalling entropy tumours are generally grade 3, whereas the low ES cell enriched, high signalling entropy tumours can be of lower grade (Fig 4.2A). This indicates that a high signalling entropy is related to but not solely determined by the level of differentiation of a tumour, and that other factors, such as inter-cellular heterogeneity may cause high signalling entropy in certain samples.

We also found that signalling entropy correlated with tumour cellularity and was highest in those samples with the greatest proportion of cancerous cells ($p < 7.7 \times 10^{-15}$). This result is consistent with our previous finding that cancerous tissue displays a higher signalling entropy its healthy counterparts [23, 293].

4.2.2.2 Signalling Entropy correlates with levels of tumour stemness in lung adenocarinoma, and is elevated in squamous-cell carcinoma We next considered three NSCLC microarray data sets: The Director's Challenge dataset profiling 398

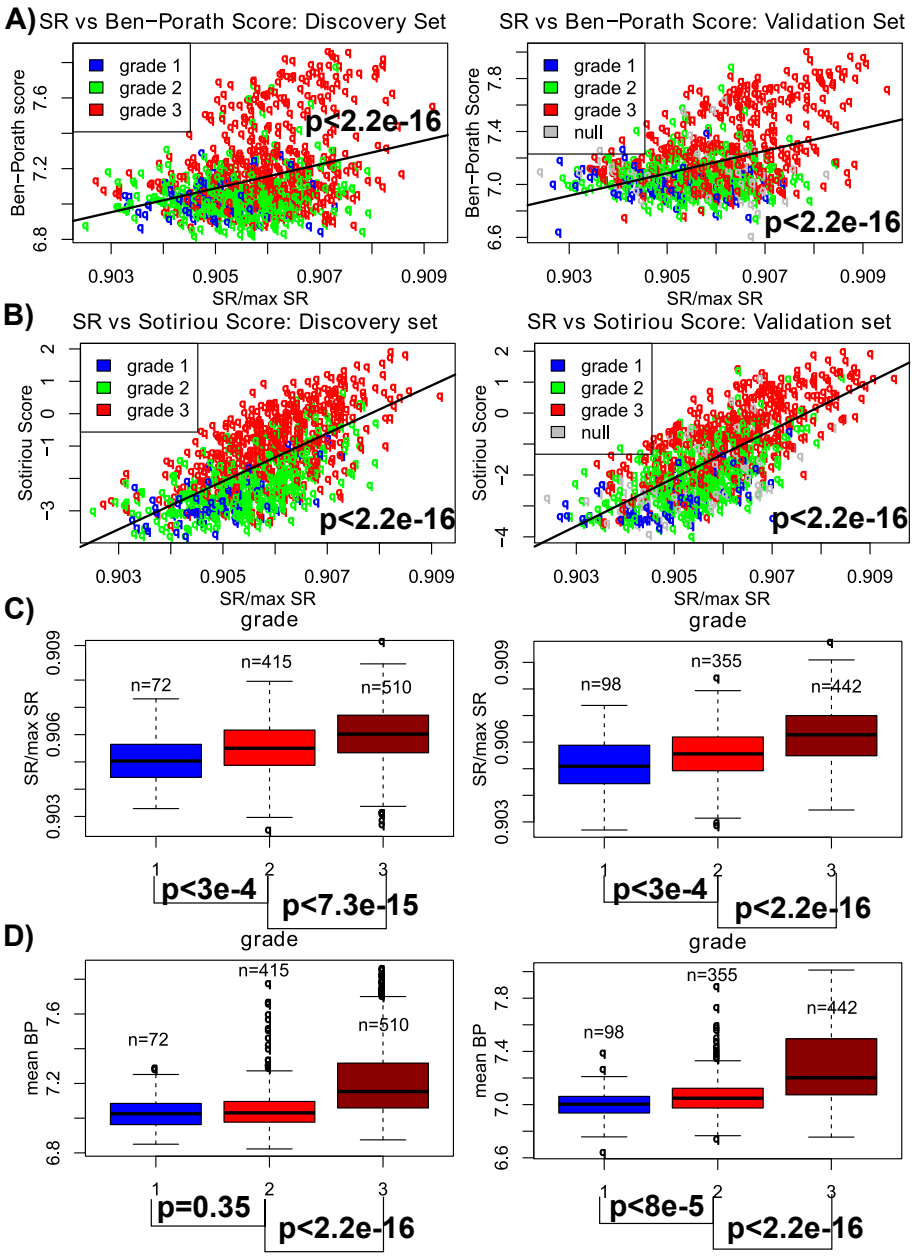

Figure 4.2: **Signalling entropy is correlated with measures of tumour stemness in breast cancer.** Signalling entropy correlates with the (A) Ben-Porath *et al.* ES cell and (B) Sotiriou *et al.* tumour grade signatures. The p-values are for linear regression. We note a bi-modality in signalling entropy for tumours with high ES cell gene expression. (C) Signalling entropy is associated with histological tumour grade. (D) The Ben-Porath *et al.* signature cannot discriminate between grade 1 and grade 2 breast cancers in the METABRIC discovery data set, p-values are from Wilcoxon tests.

primary lung adenocarcinomas [325], GSE4573 profiling 130 SCCs [326] and GSE41271 profiling 183 adenocarcinomas, 80 SCCs and 12 lung tumours of different histology [327]. In addition, we considered RNA-seq data-sets downloaded from *The Cancer Genome Atlas* (TCGA) data base (`http://cancergenome.nih.gov/`) consisting of 455 lung adenocarcinomas and 409 SCC samples.

The molecular tumour grade signature derived by Ben-Porath *et al.* [277], has also previously been applied to measuring tumour stemness in NSCLC [323]. The signature was investigated by Hassan *et al.* and found to show association with histological grade and clinical outcome in lung adenocarcinoma but not in SCC [323].

We thus evaluated the relationship between signalling entropy and histological and molecular measures of tumour stemness, as well as the relationship between signalling entropy and histological subtype in NSCLC. We found that, as in breast cancer, signalling entropy correlated strongly with the Ben-Porath *et al.* signature in both lung adenocarcinoma and SCC ($p < 2.2 \times 10^{-16}$, Fig 4.3A). Moreover, we found that signalling entropy correlated strongly with histological assessments of tumour differentiation in lung adenocarcinoma, being highest in poorly differentiated tumours, then moderately differentiated tumours ($p < 1.4 \times 10^{-4}$ poorly *vs.* moderately differentiated) and lowest in well differentiated tumours ($p < 2.5 \times 10^{-5}$ moderately *vs.* well differentiated, Fig 4.3B). In line with the findings of Hassan *et al.*, however, we found that signalling entropy, like the Ben Porath *et al.* signature, showed no association with histological grade assessments in SCC ($p = 0.17$ poorly *vs.* well/moderately differentiated).

To understand this difference we investigated the effect of histological subtype on both signalling entropy and the Ben-Porath *et al.* signature, in data sets which profiled a large number of both adenocarcinomas and SCCs. Intriguingly, we found that both signalling entropy and the Ben-Porath *et al.* signature were significantly higher in SCC as compared to adenocarcinoma ($p < 1.3 \times 10^{-10}$, Fig 4.3C). This elevation of both signalling entropy and the expression of stem cell genes in SCC, suggests that these factors may be driving the treatment resistance and heterogeneous expression of prognostic markers in this subtype [328, 329]. The fact that neither the Ben-Porath *et al.* signature nor signalling entropy correlate with histological grade in SCC may be reflective a considerable intra-tumour heterogeneity, endemic of this subtype, which imposes a lower bound on these measures and limits their variability and thus capacity to associate with histological grade assessments. Indeed, in the large TCGA data set the variance of signalling entropy was significantly higher among lung adenocarcinomas as compared to SCCs ($p = 7.6 \times 10^{-6}$).

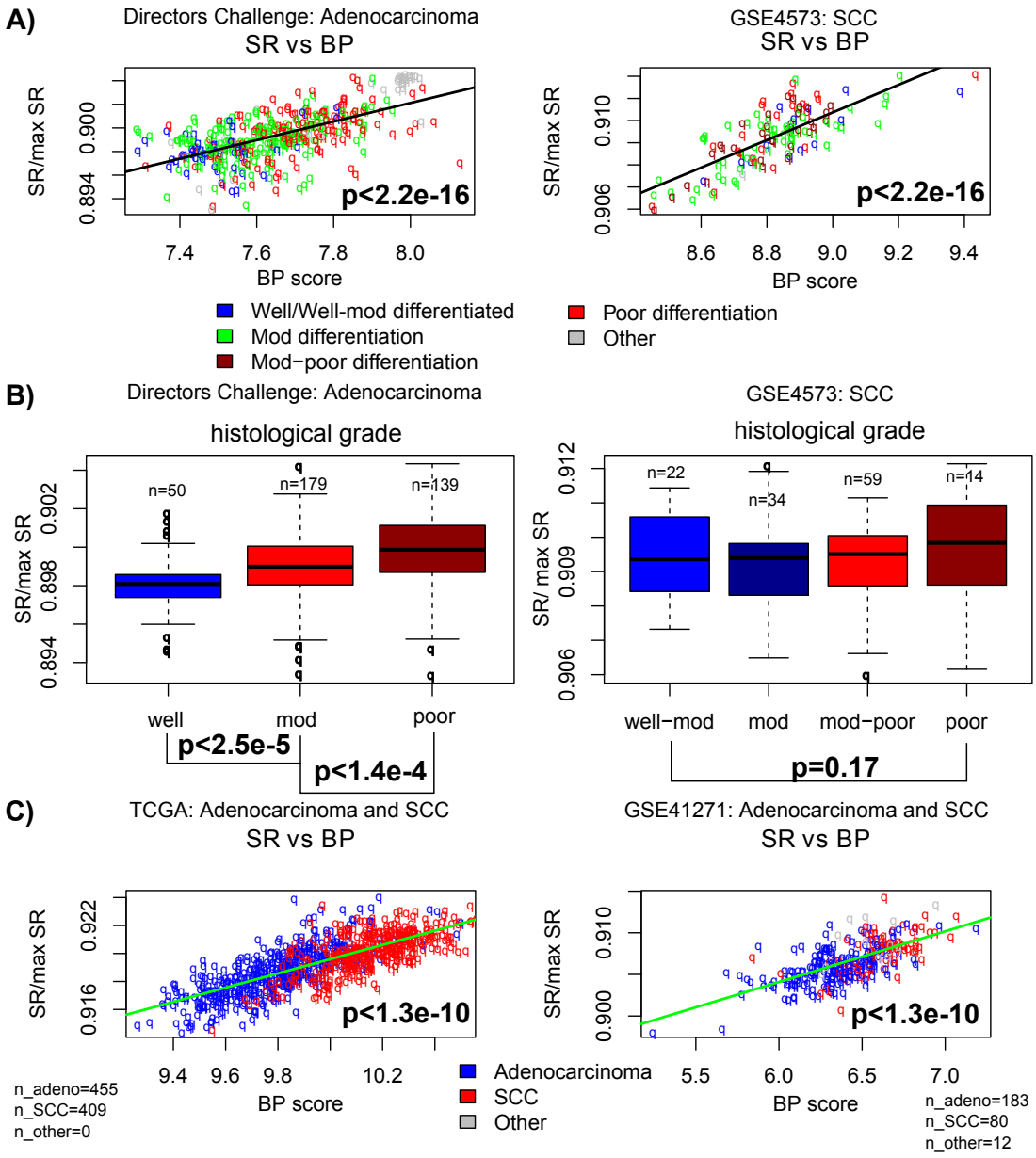

Figure 4.3: **Signalling entropy is correlated with measures of tumour stemness in lung adenocarcinoma.** (A) Signalling entropy correlates with the Ben-Porath *et al.* signature in both lung adenocarcinoma and SCC. The p-value is for linear regression. (B) Signalling entropy is only associated with histological tumour grade in lung adenocarcinoma not SCC. (C) Both signalling entropy and the Ben-Porath *et al.* signature are elevated in SCC as compared with adenocarcinoma. p-values for (B) and (C) are for Wilcoxon tests.

4.2.2.3 Signalling Entropy associates with measures of tumour stemness in prostate cancer and glioma Breast and lung cancer are our main malignancies of interest in this study. However, as a final validation that signalling entropy correlates with measures of tumour stemness across multiple malignancies, we investigated cancers arising from two more tissues of origin, namely prostate cancer and glioma.

We first considered a microarray data set (GSE6606) profiling 65 primary prostate tumours annotated with Gleason grade. A suitable molecular signature for tumour stemness in prostate cancer was derived by Penney *et al.* [330] and we found that signalling entropy significantly correlated with this signature across all prostate cancer samples ($p = 1.3 \times 10^{-4}$, Fig 4.4A). Moreover, we found that signalling entropy was elevated in high Gleason grade (≥ 8) as compared to low Gleason grade (≤ 6) tumours ($p < 0.037$, Fig 4.4B).

We next considered a microarray data set profiling 180 primary glioma tumours of differing histological grade (GSE4290). As with other malignancies we found that signalling entropy was significantly elevated in grade 3 as opposed grade 2 oligodendrogliomas and astrocytomas ($p < 0.022$, Fig 4.4C) and was significantly higher in grade 4 glioblastomas as opposed to lower grade gliomas ($p < 0.003$ Fig 4.4D).

4.2.3 Signalling Entropy associates with key clinical variables in breast and lung cancer: novel insights into pathology

Signalling entropy is thus a correlate of tumour stemness in many malignancies, though not in lung SCC, where it displays very high values. We thus postulated that other cancer clinical features may drive our measure, most likely through an association with cell potency independent intra-tumour heterogeneity. Given the extensive clinical annotation of the breast and lung cancer data sets considered, we investigated this postulate in the context of these two malignancies.

4.2.3.1 Luminal B breast cancer displays the highest signalling entropy, yet a low tumour stemness, suggesting high intra-tumour heterogeneity We first examined whether signalling entropy was associated with breast cancer subtypes.

We found that signalling entropy strongly associated with the Pfam50 intrinsic subtype classifications [136] in the METABRIC discovery and validation data sets. Normal tumours displayed the lowest signalling entropy, followed by luminal A tumours (luminal A *vs* normal $p < 2.5 \times 10^{-8}$), then *HER2* positive tumours (*HER2 vs* luminal A $p < 1.9 \times 10^{-12}$), and lastly luminal B and basal tumours (luminal B *vs HER2* $p < 0.02$, basal *vs* luminal B $p = 0.3$), which displayed statistically equivalent signalling entropies.

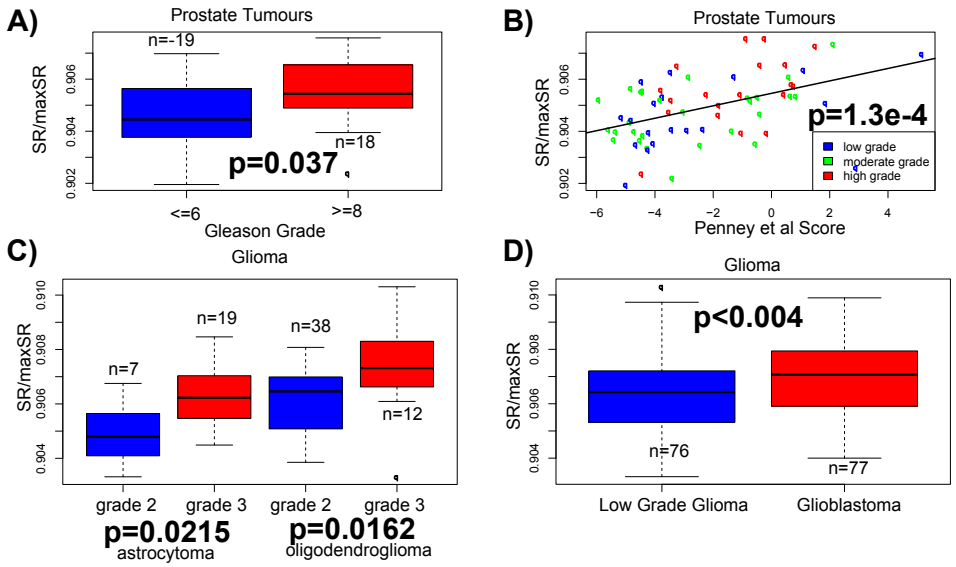

Figure 4.4: **Signalling entropy is correlated with measures of tumour stemness in prostate cancer and glioma.** (A) Signalling entropy correlates with the Penney *et al.* Gleason grade signature in prostate cancer. (B) Signalling entropy is elevated in high as opposed to low Gleason grade tumours. (C) Signalling entropy is elevated in grade 3 compared to grade 2 oligodendrogliomas and astrocytomas. (D) Signalling entropy is higher in grade 4 glioblastoma as compared to lower grade gliomas. p-values are for Wilcoxon tests and linear regression.

This result is in concordance with the report by Ben-Porath *et al.* [277], that basal breast cancers displayed a stronger enrichment of an ES cell transcriptomic signature than luminal A breast cancer (Fig 4.5).

To more fully compare the Ben-Porath *et al.* signature with signalling entropy, we examined its ability to discriminate between the intrinsic subtypes of breast cancer. As reported in the presenting paper, the signature is most enriched in the basal breast cancer subtype and least in luminal A. However the overall ordering of the intrinsic subtypes by Ben-Porath *et al.* signature enrichment is quite dissimilar from the ordering by signalling entropy. Notably luminal B breast cancers are significantly less enriched for the Ben-Porath *et al.* signature than basal ($p < 2.2 \times 10^{-16}$), despite these subtypes sharing the highest signalling entropy values.

This result is intriguing as it implies that the severity of luminal B breast cancer may be linked to a plastic signalling regime, rather than the enrichment of genes typically overexpressed in ES cells. One may conclude from this that the high signalling entropy of the luminal B subtype may be driven by intra-tumour heterogeneity rather than stemness.

Criticism of molecular subtyping has derived from the lack of diversity in the histological subtype of tumours used to define the classifications [108]. Consequently, we also examined the association between signalling entropy and histological subtype. This revealed that medullary carcinoma has the highest signalling entropy, consistent with this cancer generally being of higher grade and displaying a basal phenotype [331]. Invasive ductal carcinoma of no special type (IDC-NST) displayed the second highest signalling entropy, significantly higher than invasive lobular carcinoma ($p < 2.3 \times 10^{-5}$) mixed ductal and lobular carcinoma ($p < 1 \times 10^{-3}$) and tubular carcinoma, which all displayed a statistically equivalent signalling entropy (Fig 4.6). This result suggests that the better prognosis of tubular carcinomas compared to IDC-NST may be related to a less heterogeneous signalling regime.

We also found that signalling entropy could discriminate grade matched ER positive tumours from ER negative tumours ($p < 0.01$). Given the finding by Ben-Porath *et al.* that ER negative tumours were enriched for an ES cell signature [277], this result suggests that a stem cell-like signalling regime is more prevalent in ER negative tumours, regardless of histological grade.

Tumours carrying a mutation in p53 have also been demonstrated to be enriched for an ES cell gene expression signature [332]. In line with this result we found that grade matched p53 mutated tumours displayed a higher signalling entropy as compared their wild type counterparts ($p = 0.005$). This result is consistent with the hypothesis that p53 mutations in breast cancer can facilitate de-differentiation [332].

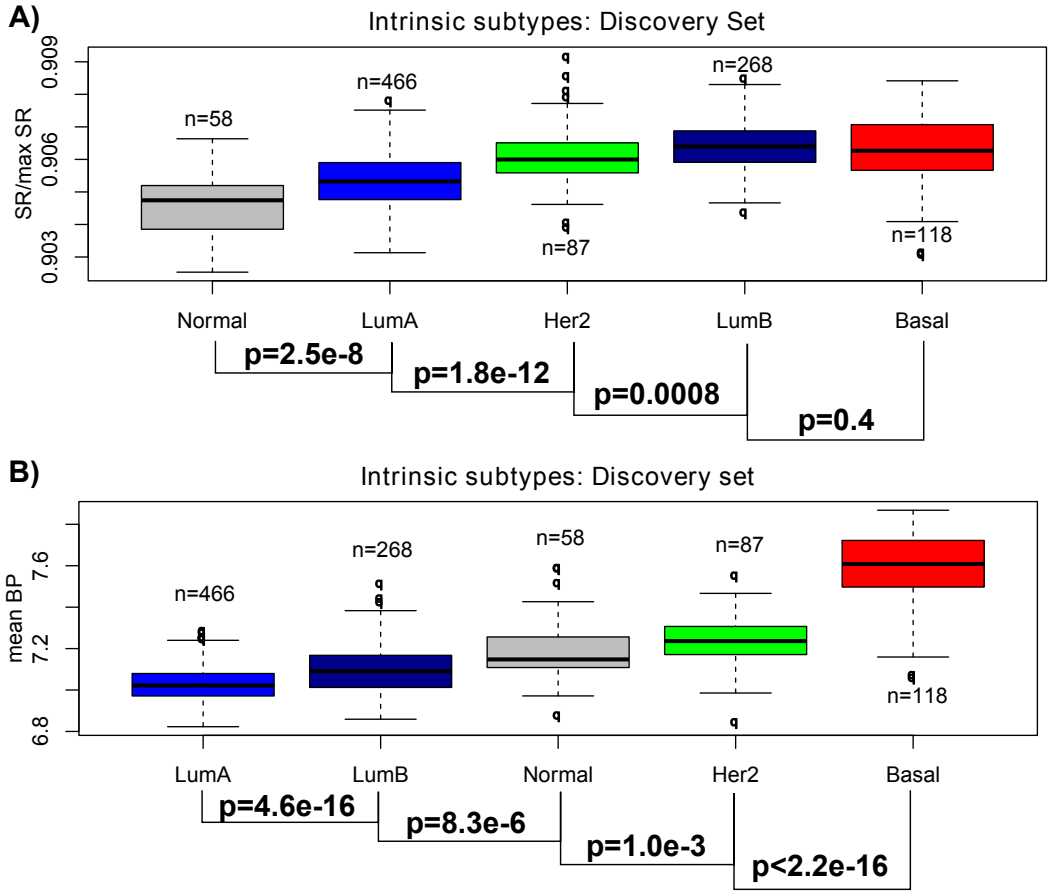

Figure 4.5: **Association between signalling entropy and the Ben-Porath *et al.* ES cell signature and intrinsic subtypes in the METABRIC dataset.** (A) Signalling Entropy is highest in the luminal B and basal subtypes. (B) Expression of the Ben-Porath *et al.* ES cell signature is highest in the basal subtype, but low in luminal B. We note that the results shown are for the METABRIC discovery data set, the results in the validation set are near identical. p-values denote Wilcoxon tests.

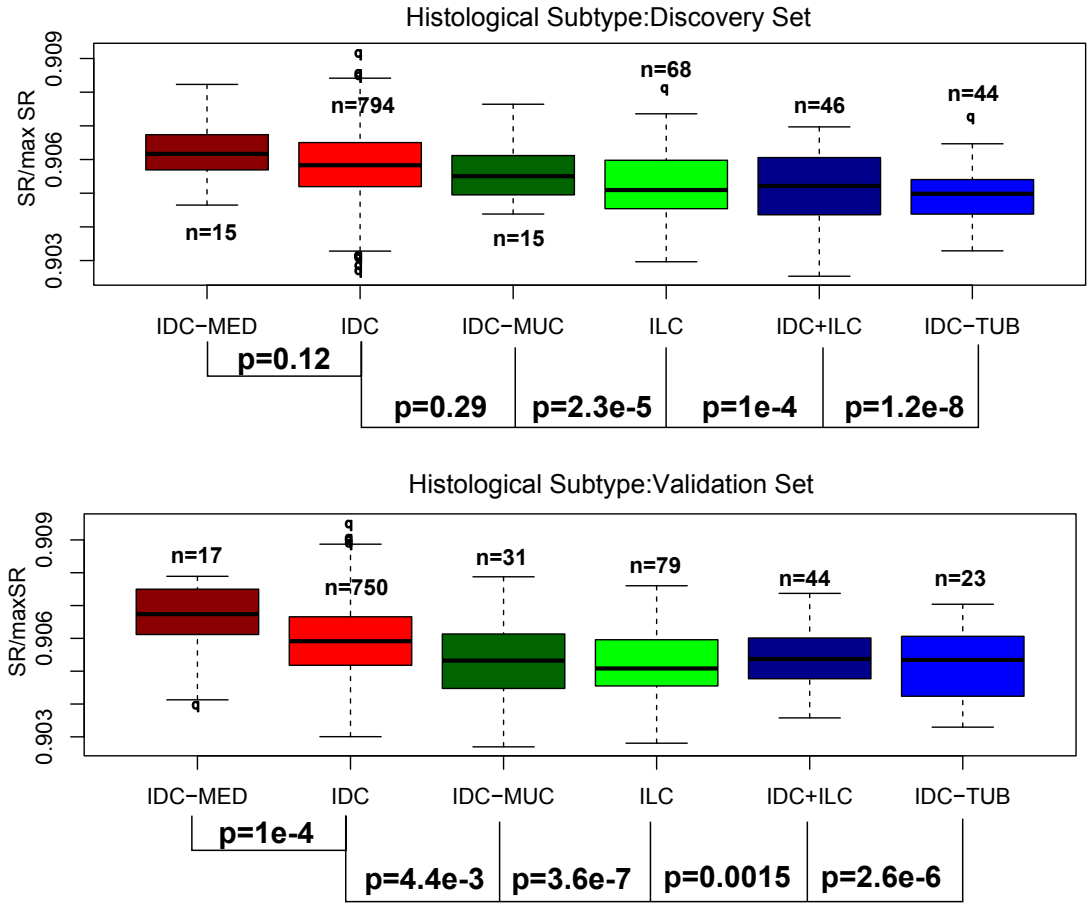

Figure 4.6: **Association between signalling entropy and histological subtypes in the discovery and validation sets of METABRIC.** IDC denotes invasive ductal carcinoma of no special type, ILC, denotes invasive lobular carcinoma, IDC-MED denotes medullary carcinoma, IDC-MUC denotes mucinous carcinoma, IDC-TUB denotes tubular carcinoma and IDC+ILC denotes mixed ductal and lobular carcinoma. p-values are for Wilcoxon tests comparing IDC with the other histological types.

4.2.3.2 Signalling Entropy associates with smoking history in NSCLC We next examined whether signalling entropy associated with different clinical variables in NSCLC.

We demonstrated earlier that SCC has a higher signalling entropy than lung adenocarcinoma and it is important to investigate why this may be. In addition to contrasting treatment resistance and prognostic gene expression profiles, the epidemiological differences between these two NSCLC histological subtypes is well studied [322, 333]. Though genetic risk factors have been identified [334], the primary cause of NSCLC is smoking. Over the last 20 years, smoking habits have changed from favouring high tar, unfiltered cigarettes towards a preference for lower tar, filtered cigarettes [333]. Concurrent with this habitual change, SCC has reduced in prevalence relative to lung adenocarcinoma and it is widely accepted that the former change caused the latter [322, 333]. In addition, patients with SCC generally have a greater exposure to cigarette smoking than those with lung adenocarcinoma. We thus postulated that signalling entropy may associate with a positive smoking history.

Hassan *et al.*, previously reported that lung adenocarcinoma patients with a positive smoking history displayed an increased expression of stem cell genes, but no association between smoking history and stem cell gene expression in SCC could be demonstrated [323]. We found, however, a strong association between signalling entropy and smoking history in both lung adenocarcinoma and SCC, manifesting by patients with a positive smoking history displaying a significantly higher signalling entropy ($p < 0.03$, Fig 4.7A-B). Moreover, unlike signalling entropy, the Ben-Porath *et al.* stem cell signature was unable to discriminate between SCC patients who were current smokers and those who were reformed smokers ($p > 0.2$, Fig 4.7C).

After confirming that signalling entropy was elevated in patients with a positive smoking history in an independent data set of mixed histologies ($p < 10^{-4}$, Fig 4.7D), we next investigated whether smoking increased signalling entropy in a dose dependent manner. We found a significant positive correlation between signalling entropy and the number of packs of cigarettes smoked by a patient per year (Pearson's $r = 0.11$, $p = 0.003$, Fig 4.7E), confirming this postulate.

These results suggest that cigarette smoking induces an increased molecular heterogeneity in NSCLC tumours, regardless of histological subtype. Moreover, this change is accompanied by the increased expression of stem cell genes in lung adenocarcinoma, but not in SCC. Consequently, one may deduce from these results that cigarette smoking

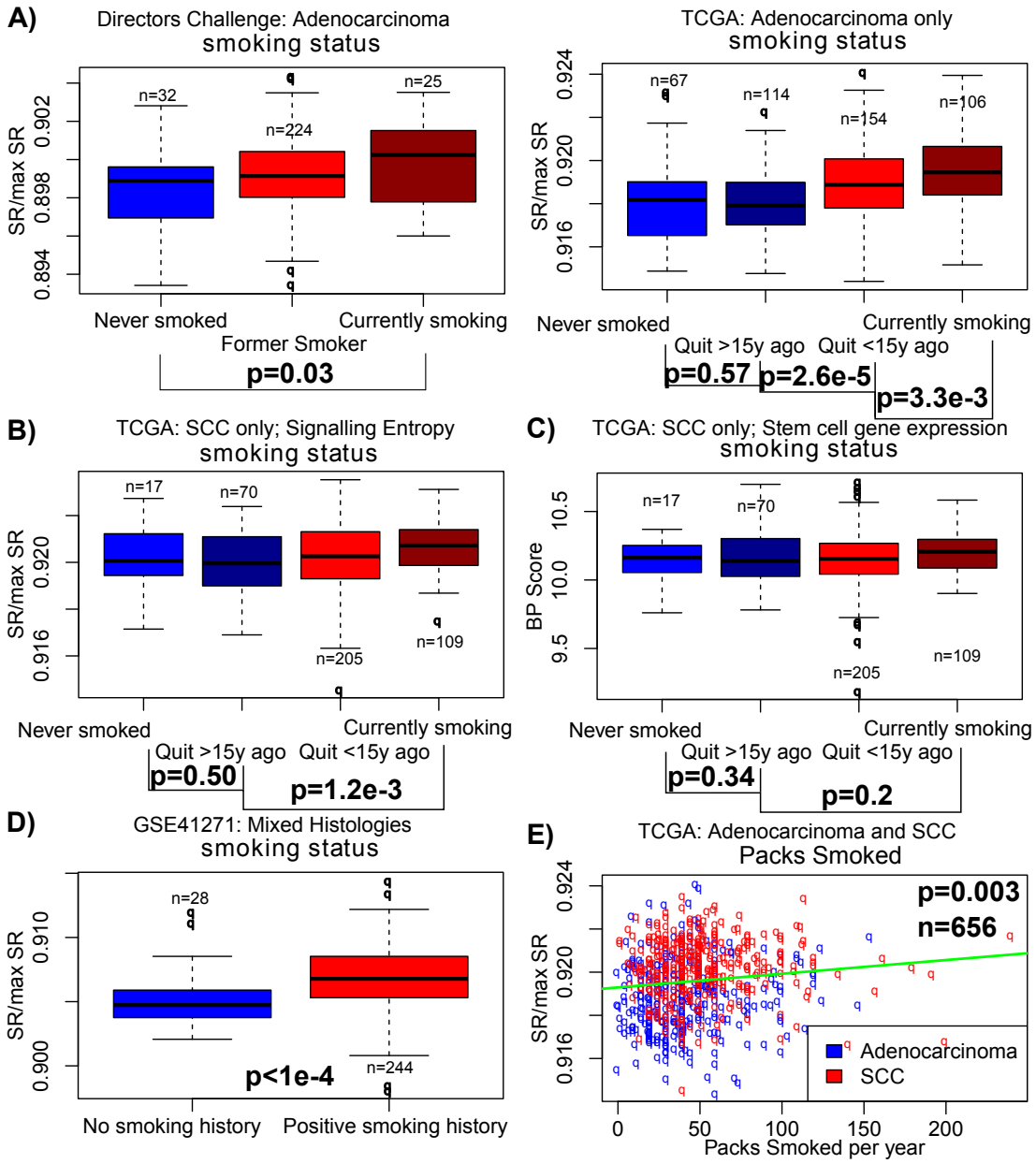

Figure 4.7: **Signalling entropy is elevated in NSCLC patients with a history of smoking.** (A) Signalling entropy correlates with smoking history in lung adenocarcinoma. (B) Signalling entropy correlates with smoking history in SCC. (C) The Ben-Porath signature doesn't correlate with smoking history in SCC. (D) Signalling entropy correlates with a positive smoking history in a data set of mixed histologies. (E) Signalling entropy positively correlates with number of packs smoked per year. p-values are for Wilcoxon tests and linear regression.

increases heterogeneity in pathway activation in lung adenocarcinoma via increasing the stemness of a the tumour, whilst in SCC where tumour stemness is high to begin with, smoking may increase signalling entropy by promoting intra-tumour heterogeneity.
In addition to the associations between signalling entropy, smoking history and stemness, we note that our measure also correlated with tumour stage in NSCLC ($p < 5 \times 10^{-5}$), was elevated in males (who typically have poorer survival) ($p < 0.042$), and was higher in patients lacking an *EGFR* mutation ($p < 4.2 \times 10^{-3}$).

4.2.4 Signalling Entropy correlates with clinical outcome in breast cancer and lung adenocarcinoma

Given the association between signalling entropy and prognostic factors, we next investigated the potential of our measure as a prognostic indicator, in a meta-analysis over a large number of breast and lung adenocarcinoma datasets.

4.2.4.1 Signalling entropy is prognostic in the major subtypes of breast cancer In order to assess the prognostic significance of signalling entropy in breast cancer, we first considered the METABRIC dataset [157]. Using outcome as a binary phenotype, we observed that patients who died of breast cancer had a higher signalling entropy than patients who were alive at last follow up; a result which was seen in both METABRIC subsets ($p < 1 \times 10^{-7}$). Using a Cox proportional hazards model, on 5 year censored survival data, we ascertained that high signalling entropy is associated with increased risk of death in breast cancer (c-index= 0.6, $p < 1.1 \times 10^{-6}$). Stratifying patients into 3 groups, representing the 3 tertiles of the signalling entropy distribution, revealed that tumours with a high signalling entropy exhibited a doubling of the hazard rate compared to low signalling entropy tumours.
As mentioned earlier signalling entropy is associated with tumour grade and ER status, however, its prognostic power is independent of these variables, as well as of stage, p53 status, tumour size and lymph node status. In addition, signalling entropy was also found to be independent of the prognostic Ben-Porath *et al.* signature [277] and the prognostic Sotiriou *et al.* grade signature [324]. Importantly, signalling entropy was significantly prognostic within each tumour grade strata; notably it was prognostic within the grade 2 stratum in both METABRIC data sets ($p < 0.036$), an important result given the difficulty in deciding treatment courses in this intermediate prognosis group [324]. The fact that signalling entropy is prognostic independently of all other measures of cell anaplasia, suggests that our measure may be capturing more than just the stemness of a tumour sample, and that intra-tumour heterogeneity may be contributing to its prognostic power.

A recent study by Venet *et al.* described prognostic associations for a number of random gene expression signatures in breast cancer [335]. To ascertain whether random effects may be driving our findings, we evaluated the prognostic associations of the three random gene expression signatures described by Venet *et al.*. We found that only one was prognostic in both discovery and validation data sets of METABRIC and that its prognostic power was determined entirely by ER status. To further assess the impact of random effects and the importance of our network, we randomised the gene expression profiles of the METABRIC data sets over the network. Performing 5 randomisations and recomputing signalling entropy for all 1980 samples in both METABRIC data sets, revealed that randomised signalling entropy did not display robust prognostic associations independently of ER status. We are therefore confident that the prognostic power of signalling entropy is not driven by random effects.

To validate the prognostic impact of signalling entropy we considered eight further independent breast cancer data sets. All these datasets described both ER positive and negative tumours with accompanying clinical outcome, profiled on either Affymetrix or Illummina platforms and totalling 1688 samples [336, 337, 338, 339, 340, 341, 342, 343], (Table 1). Meta-analysis revealed that signalling entropy is prognostic across both ER positive and ER negative samples (ER positive: c-index=0.63, 95% CI=$(0.604, 0.657)$, $p = 8.5 \times 10^{-15}$, ER negative: c-index=0.57, 95% CI=$(0.538, 0.602)$, $p = 0.032$, Fig 4.8). Five of the additional eight data sets also described histological tumour grade for each sample, allowing us to further confirm that signalling entropy is prognostic within the grade 2 stratum (c-index=0.63, 95% CI=$(0.581, 0.675)$, $p = 1.05 \times 10^{-6}$, Fig 4.8).

These results are in contrast to the performance of MammaPrint, a microarray based breast cancer prognostic signature currently being assessed in the MINDACT trial [344]. In a meta-analysis over the 10 breast cancer data sets we found that unlike signalling entropy MammaPrint was not significantly prognostic over ER negative samples (Fig 4.9).

Another popular breast cancer prognostic assay in clinical trials is OncotypeDX, which uses qPCR to quantify the expression of genes associated with survival [158]. Due to differences in the normalisation of qPCR and microarrays, a direct comparison between our measure and OncotypeDX is difficult to perform. Moreover, not all the genes required for computing the OncotypeDX recurrence score were present in all the array platforms considered. However, using a microarray version of OncotypeDX, we found that it performed comparably to signalling entropy across both ER positive and ER negative samples (OncotypeDX *vs.* signalling entropy $p > 0.13$, Fig 4.10).

Thus signalling entropy is prognostic in the major clinical subtypes of breast cancer and hence is a more robust prognostic indicator than MammaPrint.

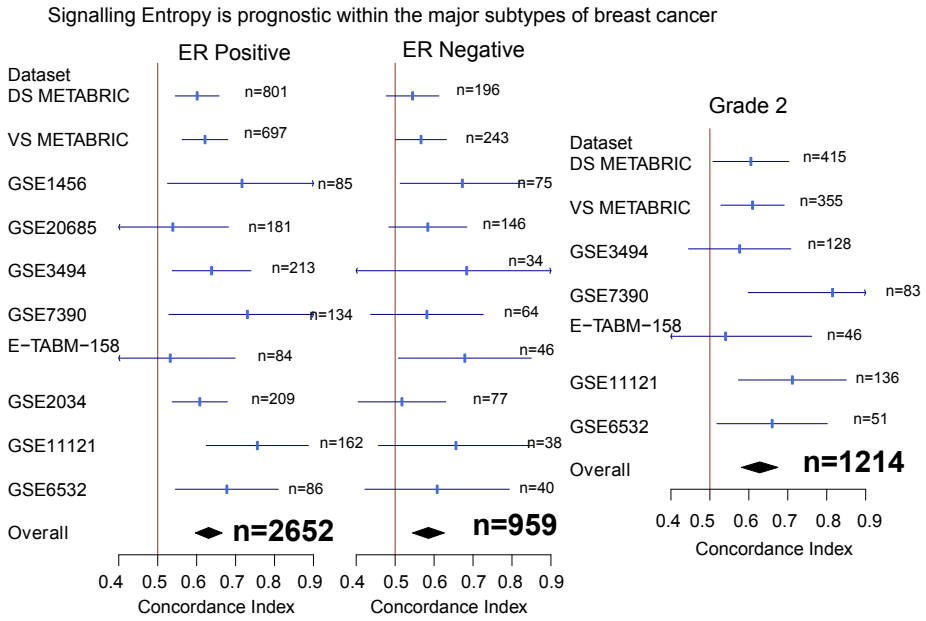

Figure 4.8: **Meta-analysis of prognostic implications of signalling entropy in breast cancer** The plots display the concordance index for signalling entropy in each data set alongside its 95% confidence interval. The overall concordance index was derived via meta-analysis using a random effects model. The vertical line denotes concordance index= 0.5, data sets where the confidence interval for the concordance index crosses this line did not reach significance. Meta-analysis of signalling entropy across 10 breast cancer data sets reveal that our measure is significantly prognostic across both ER positive and ER negative subtypes. Meta-analysis across 7 breast cancer data sets reveal that our measure is also significantly prognostic within the grade 2 stratum.

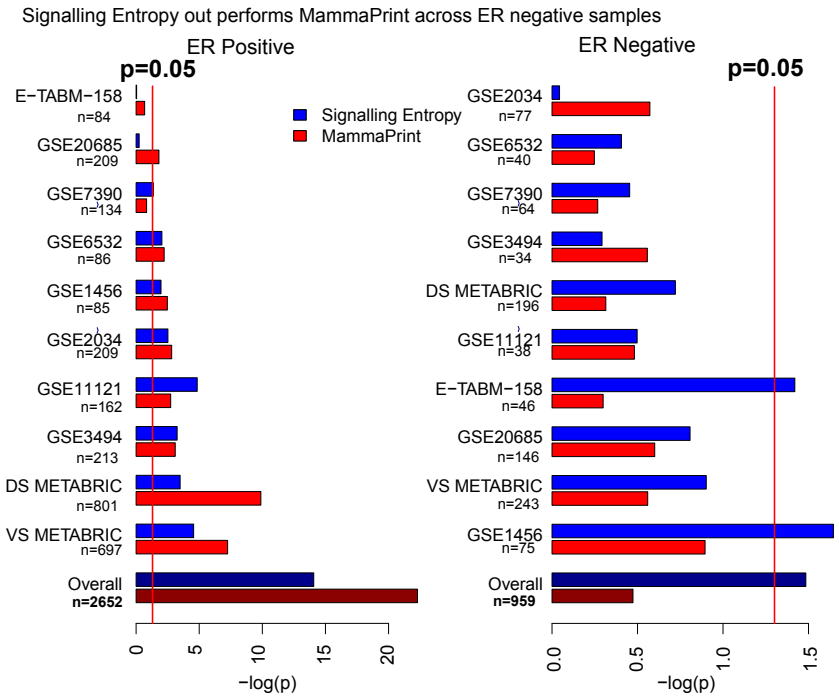

Figure 4.9: **Signalling entropy out-performs MammaPrint over ER negative samples** The plots display the negative of the $\log_{10}$ of the p-value for a survival analysis using Cox-regression on 5-year censored data, evaluating the prognostic significance of signalling entropy and MammaPrint in each data set. The overall p-value was produced by a Fisher's combined test. The vertical red line on each plot denotes $p = 0.05$; data sets in which the bar crosses this line reached significance for the corresponding score. Meta-analysis comparison of signalling entropy with MammaPrint across 10 breast cancer data sets, demonstrates that only signalling entropy is significantly prognostic across ER negative samples.

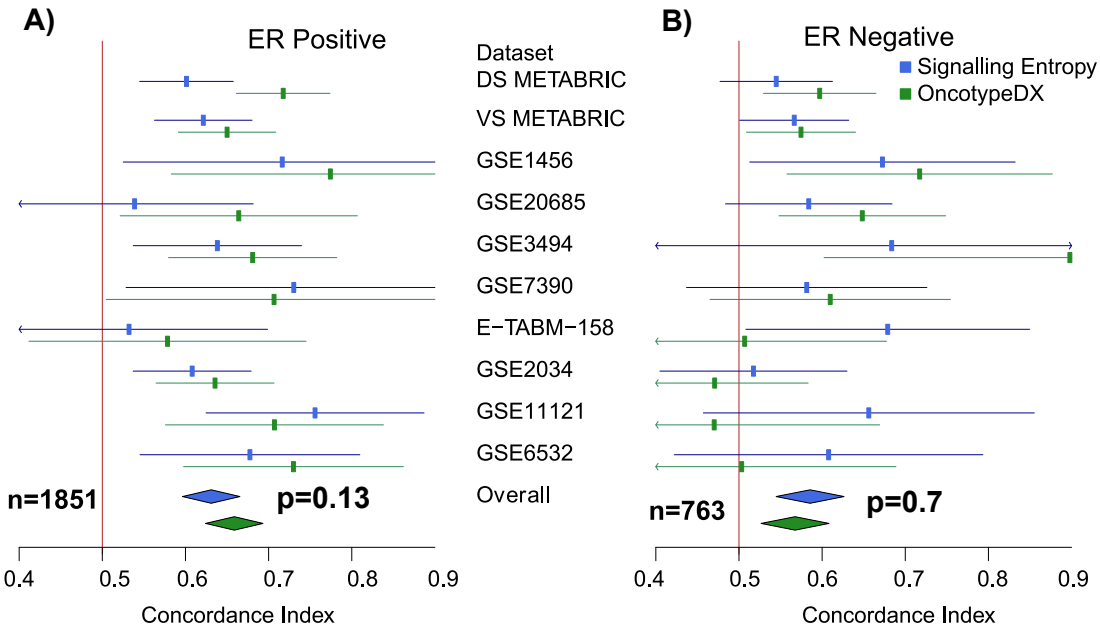

Figure 4.10: **Meta-analysis comparison of signalling entropy with OncotypeDX.** The plots display the concordance indices for signalling entropy and a microarray based approximation of OncotypeDX in each data set alongside 95% confidence intervals. The overall concordance indices were derived via meta-analysis using a random effects model. The vertical line denotes concordance index= 0.5, data sets where the confidence interval for the concordance index crosses this line did not reach significance. Meta-analysis across 10 data sets reveal that signalling entropy performs comparably to OncotypeDX across (A) ER positive samples and (B) ER negative samples.

Data Set	No. Samples	No. ER+ Samples	No. ER- Samples	No. Grade 2 Samples
METABRIC Discovery Set	997	801	196	415
METABRIC Validation Set	983	697	243	355
GSE20685	327	181	146	NA
GSE7390	198	134	64	83
GSE1456	159	85	75	NA
GSE3494	251	213	34	128
E-TABM-158	130	84	46	46
GSE2034	286	209	77	NA
GSE11121	200	162	38	136
GSE6532	137	86	40	51
Total	3668	2652	959	1214

Table 1: **Breast cancer data sets.** GEO and ArrayExpress accession numbers for the breast cancer data sets. Sample counts are provided for ER segregated and grade 2 samples.

4.2.4.2 Signalling entropy is prognostic in *stage I* lung adenocarcinoma We next investigated the prognostic power of our measure in lung adenocarcinoma. We first considered The Director's Challenge dataset profiling 398 tumours [325], and the 455 lung adenocarcinoma RNA-seq tumour samples downloaded TCGA. We found that signalling entropy was significantly lower in lung adenocarcinoma patients who were alive at last follow up as opposed to those who had died ($p < 0.03$). Fitting Cox proportional hazard models to 3 year censored data revealed that an increased signalling entropy implied a worse prognosis in lung adenocarcinoma (c-index=0.6, $p < 0.007$). We again separated patients into tertiles of the signalling entropy distribution and found that high signalling entropy conferred almost a doubling of the hazard rate, as assessed over the first 3 years following diagnosis.

As reported earlier, signalling entropy associates with tumour stage, grade and smoking status in NSCLC, importantly, however, the prognostic power of signalling entropy was independent of these clinical variables.

It is of particular note that signalling entropy is significantly prognostic if computed from either microarray or RNA-seq data sets, this result attests to the biological relevance of our measure, which is not masked by experimental technique.

To validate the prognostic power of signalling entropy in lung adenocarcinoma, we performed a meta-analysis across 4 further independent data sets consisting of a total of 522 lung adenocarcinomas (Table 2) [345, 346, 347, 327]. This revealed that signalling entropy is prognostic across all samples and across *stage I* samples (all samples: c-index=0.58, 95% CI=$(0.55, 0.60)$, $p = 1.9e \times 10^{-6}$, *stage I*: c-index=0.56, 95% CI=$(0.52, 0.60)$, $p = 0.037$, Fig 4.11 & 4.12).

Early stage lung adenocarcinoma suffers from a high relapse rate and it is important to

Data Set	No. Samples	No. *stage I* Samples
TCGA	455	243
Directors' Challenge	398	232
GSE31210	246	168
GSE37745	196	130
GSE50081	127	92
GSE41271	270	132
Total	1692	997

Table 2: **Lung cancer data sets.** GEO and ArrayExpress accession numbers for the lung cancer data sets. Sample counts are provided for *stage I* samples.

establish more robust prognostic assessments in the *stage I* subgroup for chemotherapeutic treatment stratification [348]. Sub-staging by size is currently the standard clinical approach to stratify *stage I* tumours, however, on meta-analysis we found that this stratification, unlike signalling entropy was not significantly prognostic over the *stage I* stratum (Fig 4.12).

We note that whilst no lung adnenocarcinoma gene expression based prognostic indicators are currently in clinical trials, two scores are considered viable candidates: the expression of *CADM1* [346] and a qPCR based score derived recently by Kratz *et al.* [348]. We found that both prognostic scores performed similarly to signalling entropy in a meta-analysis setting. Though we note that as with OncotypeDX, a direct comparison with the qPCR based Kratz *et al.* score is not conclusive in a microarray/RNA-seq setting.

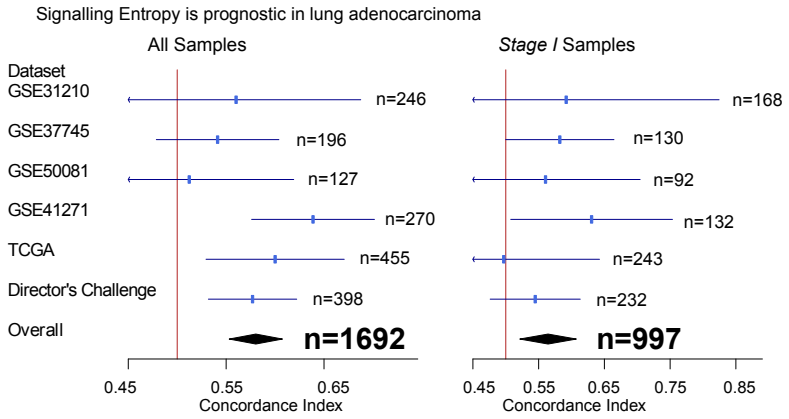

Figure 4.11: **Meta-analysis of prognostic implications of signalling entropy in lung adenocarcinoma.** The plots display the concordance index for signalling entropy in each data set alongside its 95% confidence interval. The overall concordance index was derived via meta-analysis using a random effects model. The vertical line denotes concordance index= 0.5, data sets where the confidence interval for the concordance index crosses this line did not reach significance. Meta-analysis of signalling entropy across 7 lung adenocarcinoma data sets reveal that our measure is significantly prognostic across all samples and within the *stage I* stratum.

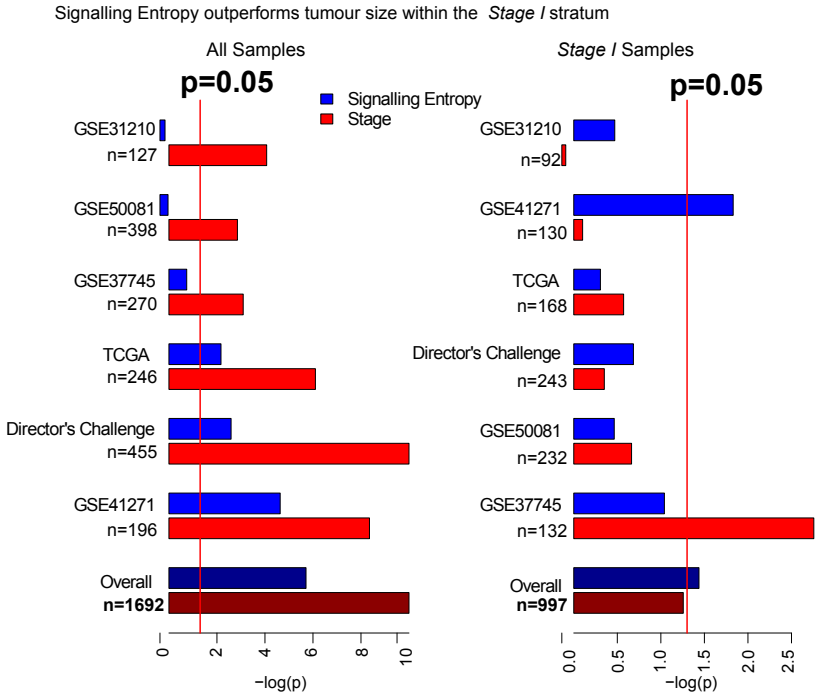

Figure 4.12: **Signalling entropy outperforms tumour size staging expression across *stage I* samples.** The plots display the negative of the $\log_{10}$ of the p-value for a survival analysis using Cox-regression on 3-year censored data, evaluating the prognostic significance of signalling entropy and tumour stage in each data set. The overall p-value was produced by a Fisher's combined test. The vertical red line on each plot denotes $p = 0.05$; data sets in which the bar crosses this line reached significance for the corresponding score. Meta-analysis comparison of signalling entropy with pathological tumour stage across 7 lung adenocarcinoma data sets, demonstrates that signalling entropy outperforms the *stage Ia/b* sub staging across *stage I* samples.

4.2.5 Signalling entropy's prognostic power in breast cancer can be represented by a small number of genes

Signalling entropy is a clear prognostic indicator in breast cancer, yet its computation requires the expression of many thousands of genes, something which is currently cumbersome and expensive for clinical application. Moreover, our measure associates with tumour grade and ER status in breast cancer and thus the factors driving its prognostic power independently of these variables is unclear. We posited that the prognostic power of our measure, independent of ER status and grade may be captured by the expression of a small number of genes.

We thus utilised correlations and prognostic associations to identify a small set of 81 genes from the METABRIC discovery data set, associated with signalling entropy's prognostic power independently of ER status and grade (Materials and Methods, Chapter 4). A *Signalling Entropy prognostic score* (SE score) was then defined from the expression of these 81 genes to ascertain whether our hypothesis was correct (Materials and Methods, Chapter 4).

Meta-analysis across the remaining 9 independent breast cancer data sets revealed that like signalling entropy, the SE score is prognostic across both ER positive and ER negative samples (ER positive: c-index=0.63, 95% CI=$(0.59, 0.67)$, $p = 4.6 \times 10^{-15}$, ER negative: c-index=0.62, 95% CI=$(0.58, 0.66)$, $p = 8.1 \times 10^{-8}$, (Fig 4.13A).

Moreover, meta-analysis further demonstrated that the SE score performed comparably to MammaPrint over ER positive samples ($p = 0.18$, Fig 4.13B), and out-performed MammaPrint over ER negative samples ($p = 0.04$, Fig 4.13C). Whence the prognostic power of signalling entropy is well captured by the expression of this small set of genes.

4.2.6 A signalling entropy based prognostic score for lung adenocarcinoma outperforms *CADM1* expression

We next investigated whether a similar SE score is capable of capturing the prognostic power of our measure via the expression of a small number of genes, in lung adenocarcinoma. Signalling entropy is correlated with, yet prognostically independent of tumour stage in lung adenocarcinoma, we therefore derived a set of 29 genes from the Director's Challenge data set, which captured the prognostic power of our measure independently of tumour stage and utilised these genes as the basis of an SE score (Materials and Methods, Chapter 4).

Meta-analysis across 5 independent validation data sets revealed that the SE score is prognostic across all samples and across *stage I* samples (all samples: c-index=0.62, 95% CI=$(0.59, 0.66)$, $p = 1.9 \times 10^{-11}$, *stage I*: c-index=0.66, 95% CI=$(0.60, 0.71)$, $p =$

A)

ER Positive				
	c-index	95% Confidence Interval		p
SE score	0.631274	0.5967612	0.6657875	4.56E-15
MammaPrint Score	0.664789	0.6302755	0.6993017	1.46E-18
ER Negative				
	c-index	95% Confidence Interval		p
SE score	0.621313	0.5804345	0.6621919	8.09E-08
MammaPrint Score	0.561133	0.5202539	0.6020113	0.02385

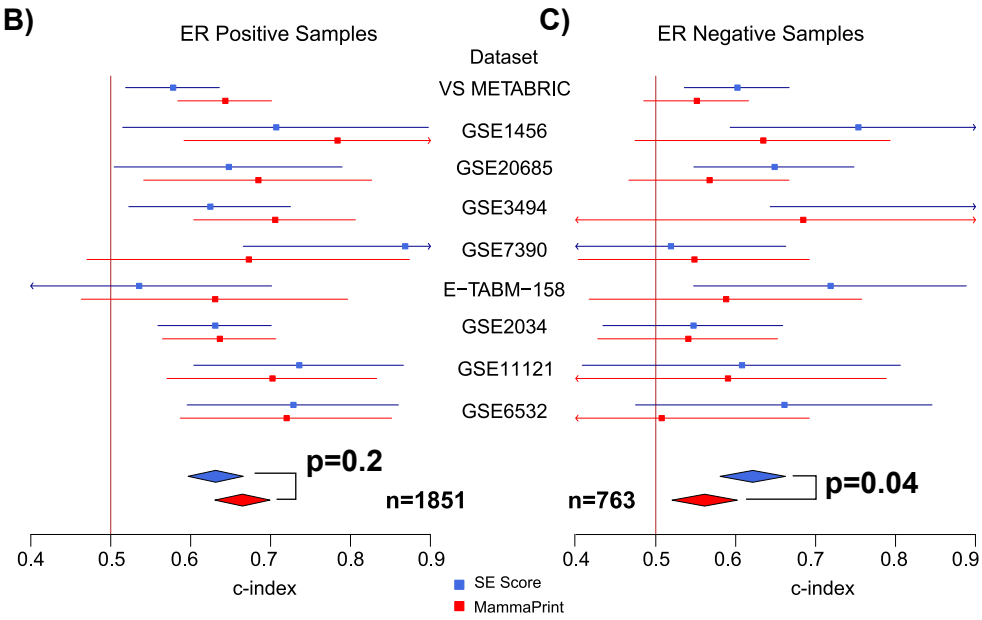

Figure 4.13: **Meta-analysis comparison of the breast cancer SE score with MammaPrint.** (A) Survival analysis statistics for the SE score and MammaPrint over ER positive samples and ER negative samples separately, c-index denotes concordance index and p denotes p-value. (B) & (C) The plots display the concordance index for the SE score and MammaPrint in each data set alongside 95% confidence intervals. The overall concordance indices were derived and compared via meta-analysis using a random effects model. The vertical line denotes concordance index= 0.5, data sets where the confidence interval for the concordance index crosses this line did not reach significance. Meta-analysis reveals that the SE score (B) performs comparably to MammaPrint in ER positive samples and (C) outperforms MammaPrint across ER negative samples. p-values are from a Fisher's combined test across all data sets.

3.35×10^{-5}, Fig 4.14).

We next compared the SE score to the leading microarray gene expression based prognostic indicator for lung adenocarcinoma, mentioned earlier, the expression of *CADM1*. *CADM1* expression was outperformed by pathological tumour stage ($p = 0.03$) in a meta-analysis. In contrast the SE score performed comparably to tumour stage ($p = 0.13$, Fig 4.14B). We note that this is in contrast to signalling entropy which performed identically to *CADM1* expression, indicating that in the generation of the SE score we have improved on signalling entropy's prognostic power.

As mentioned earlier conventional tumour sub-staging by size within the *stage I* stratum, is established clinical practice, it has thus been suggested that prognostic scores should aim to provide information which complements this staging, rather than seeks to replace it [349]. We therefore evaluated whether prognostic models which combined either the SE score or *CADM1* expression with *stage Ia/b* status within the *stage I* sub group, outperformed *stage Ia/b* status alone. We found that the SE score improved over *stage Ia/b* alone in a meta-analysis across 765 *stage I* lung adenocarcinomas ($p = 0.025$), whereas *CADM1* expression made no improvement over *stage Ia/b* ($p = 0.13$, Fig 4.14C). Whence the SE score provides a stronger candidate prognostic tool than *CADM1* expression for clinical application in lung adenocarcinoma.

4.2.7 The prognostic impact of signalling entropy is associated with genes involved in CSCs and treatment resistance

Given the power of signalling entropy as a prognostic factor in both breast and lung cancer can be reduced to the expression of a small set of genes we next investigated which pathways and processes these genes were involved in.

We performed GSEA on the genes utilised to derive the SE scores (Materials and Methods, Chapter 4). The strongest enrichment was for genes associated with poor survival in lung cancer, histological grade in breast cancer and cell proliferation. In addition, considerable enrichment was found for genes down regulated by the therapeutic agent salirasib and by *EGFR* inhibitors, as well as for genes up regulated in cell lines resistant to the doxorubicin, supporting the hypothesis that signalling entropy associates with therapeutic resistance. Enrichment was also found for gene sets associated with stem cells and certain CSC pathways. Examples include, genes down-regulated by *EZH2*, a well known stem cell gene involved in the pathogenesis of several cancers and which plays a documented role in both breast and lung CSCs [350, 351, 352, 353]. The set of genes down regulated by *CTNNB1* knock-out, a critical component of the Wnt signalling pathway, posited to be

A)

All lung adenocarcinoma samples				
	c-index	95% Confidence Interval		p
SE score	0.6229508	0.588833	0.6570686	1.89E-11
CADM1	0.6074523	0.5733345	0.6415701	6.67E-09
Stage	0.6596404	0.6270956	0.6921853	5.99E-21
Stage I lung adenocarcinoma only				
	c-index	95% Confidence Interval		p
SE score+Stage	0.6584113	0.6031536	0.713669	3.35E-05
CADM1+Stage	0.6258765	0.5706188	0.6811342	0.0002253
Stage only	0.5765846	0.5307918	0.6223773	0.0007996

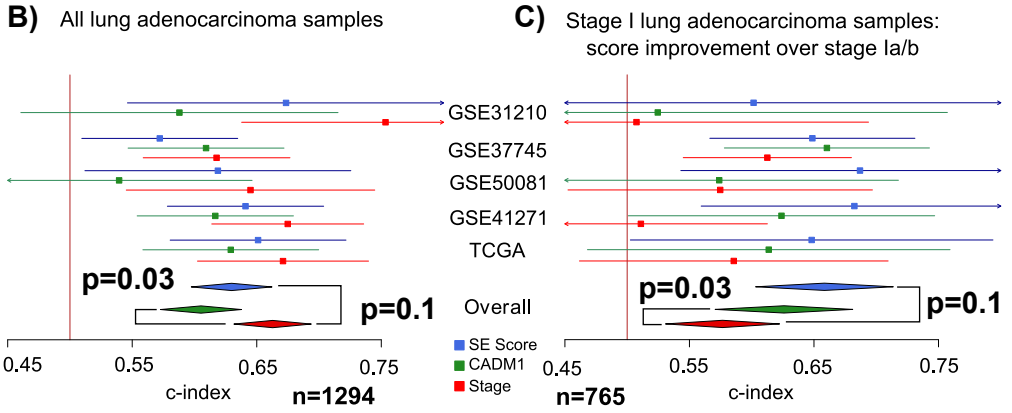

Figure 4.14: **Meta-analysis comparison of the lung cancer SE score with the expression of *CADM1*.** (A) Survival analysis statistics for the SE score and *CADM1* expression over all samples and *stage I* samples, statistics across *stage I* samples are provided for the 2 scores combined with *stage Ia/b* status, c-index denotes concordance index and p denotes p-value. (B) The plots display the concordance index for the SE score, *CADM1* expression and pathological tumour stage in each data set alongside 95% confidence intervals. The overall concordance indices were derived and compared via meta-analysis using a random effects model. The vertical line denotes concordance index= 0.5, data sets where the confidence interval for the concordance index crosses this line did not reach significance. Meta-analysis across 5 validation data sets reveal that the SE score performs comparably to tumour stage, whilst *CADM1* expression is outperformed by tumour stage. (C) The plots display the concordance index for the SE score and *CADM1* expression combined with *stage Ia/b* status, as well as *stage Ia/b* status alone, for *stage I* samples in each data set alongside 95% confidence intervals. Meta-analysis across 5 validation data sets reveal that only the SE score adds prognostic value to *stage Ia/b* status.

important in CSCs and their therapeutic resistance [354]. Targets of *BMP2* were also among the most enriched gene sets in breast but not lung cancer, which is intriguing given the role of this gene specifically in breast CSCs [355]. Enrichment was also found for many gene sets associated with immune system processes.
Thus signalling entropy is prognostically related to genes associated with both CSCs and treatment resistance, across multiple malignancies and independently of clinical variables. This result confirms our initial postulate that signalling entropy is a powerful prognostic measure, related to tumour stemness and CSCs as well as treatment resistance.

4.2.8 Signalling entropy differences between healthy and cancerous tissue correlates with tissue specific cancer mortality

As a final investigation of the prognostic power of signalling entropy across multiple malignancies, we return the Human Protein Atlas dataset considered in Chapters 2 and 3. In Chapter 3, we demonstrated that signalling entropy was elevated in the 20 cancerous tissues as compared to their 20 healthy counterparts, implying that oncogenesis induces an elevation of our measure. Moreover, different healthy and cancerous tissues display a range of signalling entropy values, and the difference between the signalling entropy of a tumour and its healthy counterpart varies across tissue types. This is indicative of some tissues generating tumours via a larger increase in signalling entropy than others.
Cancers arising in different tissues suffer from differing mortality rates, and many diverse factors are believed to underlie this [356]. Given the prognostic power of signalling entropy across multiple malignancies, however, we hypothesised that tissues forming tumours with much higher signalling entropy than their healthy counterparts, may be intrinsically more aggressive.
We obtained mortality rate data for 16 of the 20 cancerous tissues described in the Human Protein Atlas data set [356] and found that the difference in signalling entropy between cancerous and healthy tissue was significantly correlated with tissue specific mortality rate (Pearson's $r = 0.51$, $p < 0.05$, Fig 4.15).
Thus the signalling entropy increase induced by oncogenesis in a given tissue, correlates with overall tissue specific cancer mortality.

4.3 Discussion

The aim of this chapter was to investigate signalling entropy as a prognostic measure in epithelial cancer.
We previously demonstrated that signalling entropy is elevated in CSCs compared to the tumour bulk (Chapter 3) and associates with intra-sample heterogeneity (Chapter 2).

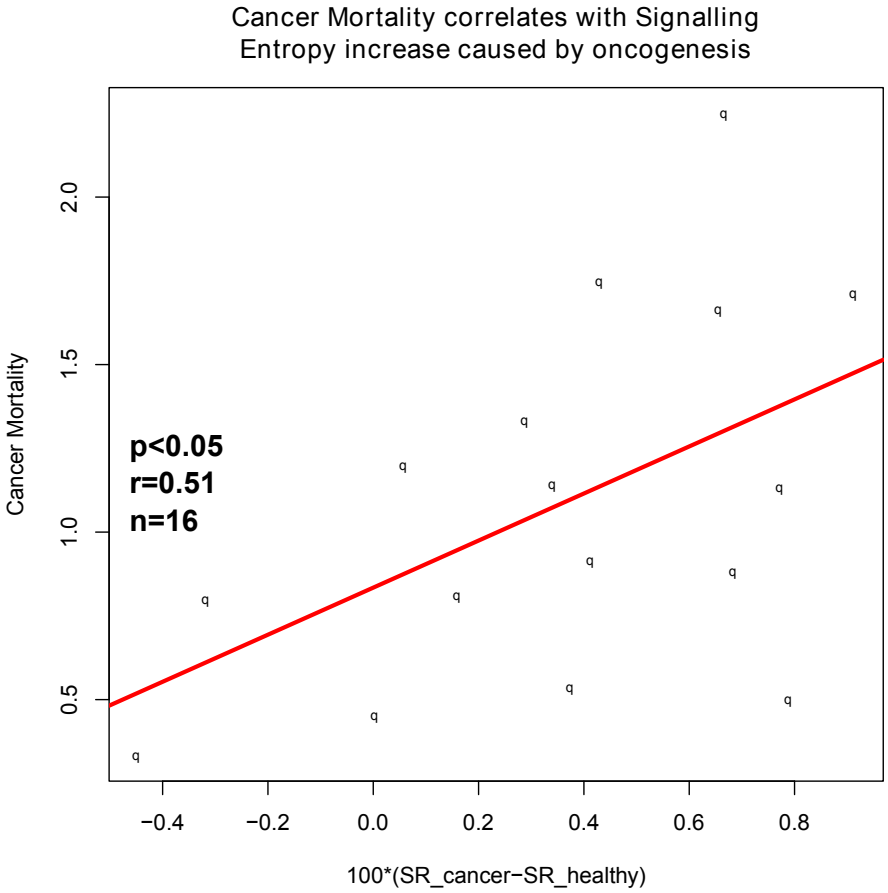

Figure 4.15: **Signalling entropy increase during oncogenesis correlates with tissue specific cancer mortality.** The signalling entropy difference between cancerous and healthy tissues is plotted against tissue specific cancer mortality rate for 16 different tissue types described by the Human Protein Atlas. We see that there is a significant positive correlation between the two variables. A p-value and Pearson's r is given for a linear regression.

As explored in Chapter 1, the discovery that CSCs are resistant to conventional therapy necessitates an evaluation of their prognostic value, and the development of targeted therapies [165, 155]. Recent observations have also demonstrated the importance of characterising intra-tumour heterogeneity in the prognostic assessment of epithelial cancers [311]. The measurement of both CSC abundance and intra-tumour heterogeneity in a clinically relevant manner, however, has presented a considerable challenge [313]. Moreover, the majority of suggested approaches are limited in sample size, and require the time consuming collection of large new data sets (such as multiple biopsies from single tumours) for validation and proof of concept.

Signalling entropy, computable from genome wide gene expression data is not so encumbered. We here considered over 5000 primary tumour samples, demonstrating robustly that our measure of tumour stemness, related to CSC abundance and intra-tumour heterogeneity correlates with clinical outcome in breast cancer and lung adenocarcinoma.

As explored in Chapter 1 breast cancer is the most common cancer among women and a foremost cause of cancer death worldwide. Recently, advances in breast cancer stratification for prognostic evaluation and treatment, have been achieved by the consideration of gene expression assays such as OncotypeDX and MammaPrint [357, 158]. The power of these indicators in ER positive breast cancer has been demonstrated [358, 359], however, they ignore intra-tumour heterogeneity and CSCs and thus likely under-perform [311]. We found that signalling entropy performed comparably to these indicators across ER positive breast cancers and outperformed MammaPrint over ER negative samples (a direct comparison to OncotypeDX is difficult due to non-comparability of data types).

NSCLC is the leading cause of cancer death worldwide [356] and the response of the condition to therapy is often limited [360]. Moreover, survival benefit from adjuvant chemotherapy has only been demonstrated for *stage II-III* patients [348]. These circumstances motivate the development of more sophisticated staging assays in NSCLC, particularly within the *stage I* stratum. Two key candidates for such an assay are the Kratz *et al.* signature [348] and the expression of *CADM1* [346]. Both scores, however, have been considered limited due to lack of consideration of micro-environmental and CSC influences on prognosis [361]. We found that signalling entropy performed comparably to these indicators but that a SE score derived from the expression of a small number of signalling entropy related genes outperformed *CADM1* expression even across *stage I* samples (a direct comparison to the Kratz *et al.* score is difficult due to non-comparability of data types).

We note that there exist many other sophisticated prognostic signatures for breast cancer, derived from within the DREAM challenge consortium, and several of these, like

signalling entropy, have demonstrated improvement over MammaPrint or OncotypeDX [362, 363, 364, 365]. The aim of our work was to introduce a new prognostic measure of signalling promiscuity, which by approximating CSC abundance and intra-tumour heterogeneity may prove a basis by which to improve the construction of prognostic models for epithelial cancers. A direct comparison of signalling entropy to the prognostic indicators from the DREAM challenge, is thus inappropriate. However, we believe that incorporation of our measure into the construction of such indicators may yield improved results and this is a topic for further work.

We note that in addition to prognostic implications, signalling entropy also provides insight into the effect of clinical variables on tumour composition. Most notably, our results are compatible with luminal B breast cancer displaying a considerable intra-tumour heterogeneity, and with cigarette smoking inducing increased heterogeneity in lung SCC.

Finally we demonstrated that the difference between cancerous and healthy signalling entropy is correlated with tissue specific cancer mortality. Suggesting that our measure captures an intrinsic property associated with cancer aggression.

We thus propose signalling entropy as a powerful and readily applicable tool for assessing the prognostic impact of intra-tumour heterogeneity, and CSC abundance. In doing so we acheive a major goal of the thesis: the application of an entropic network theoretic tools to the understanding of malignancy.

4.4 Materials and Methods, Chapter 4

4.4.1 Expression Data

For our initial survival studies in breast cancer and for identification of signalling entropy associations with clinical variables we considered the METABRIC dataset [157] described in Chapter 2.

Eight further independent breast cancer data sets were located via Oncomine, the GEO database and ArrayExpress [366, 270, 367]. All studies considered both ER positive and ER negative breast cancers and were profiled on either Affymetrix or Illumina platforms [336, 337, 338, 339, 340, 341, 342, 343]. With the exception of two, all studies were annotated with overall survival data, in the two remaining cases, relapse and distant metastasis were used as proxies for survival. Two studies also lacked ER immunohistochemistry, in these cases, a dip test was utilised to confirm that the distribution of *ESR1* expression was significantly bimodal, before partitioning around medoids was utilised on *ESR1* expression to generate two clusters dividing samples into ER positive and ER negative subgroups. The sample counts, stratified by ER status for all data sets used are

provided in Table 1, alongside GEO and ArrayExpress accession numbers. Normalised data from each study was downloaded from the GEO database or ArrayExpress. Quantile normalisation was subsequently performed across all samples within each study.
For our investigation into NSCLC we first considered the TCGA data set of RNAseqv2 data profiling 455 lung adenocarcinomas and 401 SCCs tumours and the Director's Challenge microarray data set profiling 398 lung adenocarcinoma tumours, with accompanying clinical annotation. Processed data from both studies was quantile normalised and log transformed.
Four further independent NSCLC data sets with survival data were located via Oncomine, and the GEO database [366, 270, 347, 327, 326, 325]. The sample counts, stratified specified for *stage I* samples, for all data sets used are provided in Table 2 Normalised data was downloaded from the GEO database and quantile normalised within each study.
For each study separately probes in the microarrays and sequences in the RNAseqv2 data were matched to unique EntrezGene identifiers; probes or sequences mapping to the same identifier were averaged over.

4.4.2 Protein Expression data

The protein expression data considered was as described in the Materials and Methods, Chapter 3.

4.4.3 Construction of the PIN

The PIN used in this study was described in Chapter 2.

4.4.4 Signalling Entropy

Signalling entropy was computed as described in Chapter 2.

4.4.5 GSEA

GSEA was implemented using a Fisher's exact test to compare gene lists associated with signalling entropy's prognostic power against the gene sets defined by the Molecular Signatures Database [368, 369]. Computations were performed via software downloaded from the Molecular Signatures Data Base (`www.broadinstitute.org/msigdb`) [369, 368].

4.4.6 Transcriptomic tumour stemness signatures

We consider three transcriptomic tumour stemness signatures, namely the ES cell based 100 gene signature of Ben-Porath *et al.*, the 97 gene breast cancer grade signature of

Sotiriou *et al* and the 157 gene prostate cancer grade signature of Penney *et al.* [330]. These signatures were selected due to their similarity to signalling entropy in their clinical variable association. Like signalling entropy all signatures are prognostic and correlate with tumour stemness. The Ben-Porath *et al.* signature also correlates with pluripotency in healthy tissue and associates with ER status and molecular subtype.
All genes in the Ben-Porath *et al.* signature are positively correlated with tumour grade, and thus we compute a score for this signature as the mean expression of these genes in a sample.
The Sotiriou *et al.* and Penney *et al.* signatures contains genes both up-regulated and down-regulated in higher grade cancers compared to lower grade. We therefore compute a score for these signatures as the statistic of a t-test evaluating the hypothesis that the expression of the up-regulated genes is higher than that of the down-regulated genes in each sample.

4.4.7 The SE Score in Breast Cancer

The breast cancer SE score was computed from the METABRIC discovery data set from prognostic genes, which were correlated or anti-correlated with signalling entropy independently of grade and ER status, and whose prognostic power was also independent of grade and ER status. This gene set of 320 genes was refined by fitting a Cox proportional hazards model on 5 year censored data, using all the identified genes as covariates and deleting genes which were not significantly prognostic, independently of others in the gene set. This resulted in a small set of 81 genes, 10 of which were negatively correlated with signalling entropy and 71 of which were positively correlated. These genes are provided in a supplementary table to our publication [370]. An SE score was then defined as the t-statistic evaluating the hypothesis that the 71 positively correlated genes are expressed more highly than the 10 negatively correlated genes (after z-score normalising the data for each gene, across samples).
Before survival analysis the SE score was normalised by its standard deviation within each study.
We note that by using signalling entropy to refine a set of prognostic genes identified by Cox regression, our approach refines the feature selection approach based on correlation with outcome [357]. Consequently, the genes utilised to construct our SE score are both correlated with outcome and with signalling entropy and thus should provide a prognostic indicator representative of signalling promiscuity. Criticism of feature selection for prognostic classifiers based on gene sets ranked by correlation with outcome has stemmed from the considerable discordance of such features between data sets [371, 372]. By using

signalling entropy to refine the prognostic gene set we found that this gene set instability was reduced. The genes which were both prognostic and correlated with signalling entropy showed more concordance between discovery and validation sets of METABRIC as compared to the genes which were only prognostic. Moreover, this increase in overlap was significantly higher than would be expected by chance ($p < 10^{-5}$, based on re-sampling size matched sets of prognostic genes and assessing overlap).
To further confirm this increased robustness, we derived a set of genes for constructing an SE score from the METABRIC validation set, using an identical procedure to that performed on the discovery set. This gene list was slightly shorter than for the discovery set (55 genes, 34 positively correlated and 13 negatively correlated with signalling entropy) but had an overlap of 4 genes, significantly more than would be expected by chance ($p = 0.012$, based on re-sampling size matched sets of prognostic genes and assessing overlap).

4.4.8 The SE Score in Lung Adenocarcinoma

For the lung adenocarcinoma data the SE score was computed analogously to the computation of the breast cancer SE score, using the Director's Challenge data set as a discovery set. The only differences were that that adjustment was made for tumour stage rather than grade and ER status and survival analysis was performed on 3 year censored data rather than 5 year. The basis of the score was a small set of 27 genes, 8 of which were negatively correlated with signalling entropy and 19 of which were positively correlated. These genes are provided in a supplementary table to our publication [370]. The SE score was then computed as the statistic of the t-test evaluating the hypothesis that the genes negatively correlated with signalling entropy were less expressed than those positively correlated (after z-score normalising each gene, across samples).
Before survival analysis the SE score was normalised by its standard deviation within each study.

4.4.9 MammaPrint and *CADM1* expression

A MammaPrint score was assigned to each sample in the 10 breast cancer datasets and was evaluated from the expression of the 70 genes required to define the signature as the t-statistic comparing the genes found to positively correlate with survival against those negatively correlated with survival in the study of [357].
Before survival analysis the MammaPrint score was normalised by its standard deviation within each study.
CADM1 expression was found in all 6 lung adenocarcinoma data sets and was similarly

normalised by its standard deviation within each data set before survival analysis.

4.4.10 OncotypeDX and Kratz *et al.* score approximations

OncotypeDX and the Kratz *et al.* score are both derived from qPCR rather than from microarrays like MammaPrint and *CADM1* expression. Consequentially, the comparison of signalling entropy and SE score to these indicators is not conclusive. This is because we only consider microarray and RNA-seq data, which though correlated, is normalised differently to qPCR data. Moreover some of the array platforms considered lack the expression of certain genes required to compute the OncotypeDX and Kratz *et al.* scores. We therefore approximate these scores from our data to allow a rough comparison. In the case of OncotypeDX, a score was computed from the expression of 21 genes (or as many as were represented in the data set) in each breast cancer microarray sample using the formula defined by reference [159]. In the case of the Kratz *et al.* score, this was similarly approximated from the expression of 14 genes (or as many as were represented in the data set) in each lung adenocarcinoma microarray or RNA-seq sample, using the formula defined by reference [348].

4.4.11 Meta-analysis of prognostic scores

Meta-analysis was performed to combine survival statistics for signalling entropy, the SE score, MammaPrint, *CADM1* expression and the approximations of OncotypeDX and the Kratz *et al.* score. Concordance indices were computed for each prognostic measure in each data set considered and combined using a random effects model. The p-values were combined using Fisher's combined test. Forest plots were generated using the *survcomp* package in R [373].

4.4.12 Evaluation of random gene expression signatures

The 3 random gene expression signatures described by Venet *et al.* were obtained from the supplementary information provided in reference [335]. Following the analysis described by Venet *et al.*, probes in the METABRIC data sets mapping to genes in each random gene set were extracted, genes which were associated with cell cycling as described by Ben-Porath *et al.* [277] were removed. The expression values were then median polished and a principal component analysis was performed. Samples were then partitioned into two groups, for each score via the median PC1 value. A Cox regression on 5 year censored data was then performed to evaluate the prognostic power of each random gene expression signature in each METABRIC data set. Only KRISHNAN2007DEFEAT was significantly

prognostic in both METABRIC data sets, however partition of the datasets by ER status mitigated this prognostic association.

4.4.13 Mortality rate data

Tissue specific cancer mortality rate data for 16 of the 20 cancerous tissue types described by the Human Protein Atlas were obtained from reference [356].

5 Network Rewiring in Facioscapulohumeral Muscular Dystrophy

5.1 Introduction

In our work so far we have motivated the use of signalling entropy for the investigation of complex biological phenotypes, from both analytical and data driven perspectives. It is clear that signalling entropy associates with the differentiation potential of a sample in both healthy and pathological contexts. Our measure has also has proven powerfully prognostic in epithelial cancer (a pathology epitomised as a 'caricature of healthy development'[374]), where its relationship to intra-sample heterogeneity has provided useful insights. Given this association between signalling entropy and development it is only natural to consider our measure in the context of other developmental disorders.

Arguably a highly suitable such disorder is FSHD. As explored in Chapter 1, FSHD is the most prevalent muscular dystrophy [176], yet no treatment currently exists. The condition is characterised by asymmetric, skeletal muscle atrophy affecting specific muscle groups, often associated with features including retinal vasculature abnormalities and sensorineural hearing loss [188, 186, 187].

Genetic investigations have proposed aberrant expression of a transcription factor *DUX4* as a unifying driver of FSHD pathogenesis. High levels of *DUX4* expression result in apoptosis and oxidative stress sensitivity, whilst low levels have been shown to decrease expression of the myogenic regulatory factor *MyoD* and affect myoblast differentiation and function. Moreover, myotube formation is compromised in FSHD, with myotubes possessing either atrophic or disorganised phenotypes [235]. Whilst some studies have posited that *DUX4* inhibits differentiation [375], others have suggested that an acceleration of myogenesis in FSHD drives the aberrant myotube phenotypes [241]. Given that signalling entropy is an unbiased transcriptomic measure of cell potency, it is important to assess the effect of *DUX4* and related genes on our measure.

We here reveal that *DUX4* over-expression results in an elevated signalling entropy, sug-

gesting it prevents differentiation rather than accelerates it.

We next investigate the myogenesis programme in FSHD patient derived myoblasts in detail, to experimentally confirm inhibition of differentiation, as suggested by signalling entropy. As explored in Chapter 1, FSHD is a highly heterogeneous condition, with variable penetrance even among monozygotic twins [376]. Hence the validation of well controlled cellular models for the pathology is of value. Here we consider a human model of FSHD consisting of two immortalised cell lines isolated from a mosaic patient [241]. One cell line is healthy (54-6) and possesses 13 D4Z4 repeats, whilst the other is pathological (54-12) and possesses only 3. The lines are otherwise isogenic and thus represent a perfectly controlled model of FSHD.

This model has not previously been characterised, with respect to *DUX4* phenotypes and given the relevance of *DUX4* expression to our investigation such characterisation is important. We demonstrate that the model adheres well to results found previously in primary culture: 54-12 displays a proliferation defect, a sensitivity to oxidative stress and an atrophic myotube morphology. We also demonstrate that *DUX4* is over-expressed in 54-12, albeit at very low levels.

Following this characterisation, we next perform high throughput imaging of myogenic differentiation in the two cell lines. Morphological analysis using image processing software reveals that the *DUX4* expressing pathological line shows considerable defects at both the alignment and fusion phases of myotube formation, resulting in an overall retardation of muscle differentiation, as predicted by signalling entropy.

As mentioned earlier we found that myoblasts derived from FSHD patient muscle display very low levels of *DUX4*. Moreover, many genes and pathways have been implicated as perturbed in FSHD and by *DUX4* expression, however, the overlap between such gene sets has not been rigorously assessed by network theoretic techniques. Consequently, attribution of the FSHD phenotype to *DUX4* expression remains a challenge.

FSHD is thus a pathology which can benefit from the application of local network theoretic tools, to elucidate the relationship between genes and pathways driving the pathology and provide the basis for a prioritisation of therapeutic targets. Given the value of signalling entropy in the understanding of *DUX4* expression, we employ its local analogue InSpiRe (introduced in Chapter 2), to understand genes and pathways perturbed in FSHD.

We performed a meta-analysis with InSpiRe on multiple FSHD and healthy human control sample data sets to identify pathological network rewiring. Rewiring associated with ageing, disuse atrophy, inflammation and muscle wasting was then subtracted to construct a unified FSHD-specific disease network. We further revealed by consideration of transcriptomic data describing *DUX4* over-expression, that the expression of genes in our

FSHD network can be directly attributed to *DUX4*.

Our network confirms previous findings on FSHD molecular mechanisms such as the role of perturbed myogenesis, oxidative stress sensitivity, actin cytoskeletal signalling and p53-mediated apoptosis [186, 377, 181, 213]. Importantly, we also describe novel FSHD molecular mechanisms. Notably, local network measures revealed β-catenin at the centre of our network, as the main coordinator of FSHD signalling. To determine whether *DUX4* activates β-catenin signalling, we investigated downstream β-catenin targets in *DUX4* expressing myoblasts. This confirmed significant *DUX4*-mediated perturbation of β-catenin signalling.

Thus we here reveal that *DUX4* expression increases signalling entropy, suggesting a retardation rather than acceleration of myogenesis in FSHD. We confirm this finding experimentally by high throughput time course imaging of FSHD cellular models. We next apply the InSpiRe algorithm to a meta-analysis of FSHD, deriving a detailed map of molecular mechanisms, which we attribute directly to *DUX4* expression. Finally we experimentally confirm β-catenin signalling is perturbed in FSHD, validating the capacity of InSpiRe to detect novel pathomechanisms.

5.2 Results

5.2.1 Over-expression of *DUX4* increases signalling entropy

5.2.1.1 Overview As explored above and in Chapter 1, *DUX4* is the primary FSHD candidate gene, yet very little is known about how *DUX4* expression may lead to FSHD pathogenesis. Two models of *DUX4* expression have previously been investigated. The first is the inducible i*DUX4* C2C12 murine myoblast cell line, in which cre-mediated recombination was utilised to place the *DUX4* gene under the control of a doxycycline inducible promoter [175]. The second model is the introduction of a vector containing the *DUX4* gene into a cell of interest, resulting in high levels of expression [234]. The latter has the advantage or permitting the investigator to modify the *DUX4* gene and investigate how its sequence pertains to function.

Precisely such a strategy was employed by a previous PhD student, Paul Knopp, who generated a panel of *DUX4* constructs in the pMSCV-IRES-eGFP retroviral backbone. In addition to full length *DUX4* and the homologous *DUX4c* located centromeric to *DUX4* but lacking the C-terminal domains [218], several modified *DUX4* constructs were investigated to further probe the role of the C-terminus. These included tMAL*DUX4* [216], a putative *DUX4* splice variant initiating with the amino acids MAL and lacking the C-terminal domains. In addition, the tMAL*DUX4* construct was fused to a VP16

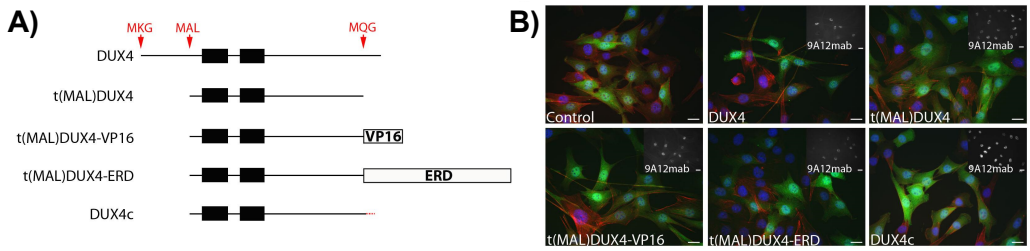

Figure 5.1: **The *DUX4* retroviral constructs introduced to murine myoblasts before transcriptomic analysis.** (A) Schematic representation of *DUX4* coding sequences: full length *DUX4*, *DUX4c* a truncated *DUX4* variant (tMAL*DUX4*) and tMAL*DUX4* C-terminal VP16 and ERD fusions, cloned into the pMSCV-IRES-eGFP retroviral expression vector. (B) Retroviral expression of *DUX4* constructs in C2C12 myoblasts immuno-stained for eGFP (green), actin (red), DAPI (blue) and with the 9A12 *DUX4* monoclonal antibody (white, in inset panel). (immunocytochemistry performed by Paul Knopp., figure adapted from Knopp *et al.*, 2015, *in preparation*).

trans-activation domain from the human herpes simplex virus 1 VP16 protein, to generate tMAL*DUX4*-VP16 a transcriptional activator of *DUX4* targets. Conversely, to repress transcriptional targets of *DUX4* via recruitment of a repressive complex, tMAL*DUX4* was fused to the N-terminus of Drosophila melanogaster engrailed (residues 2-298) to create tMAL*DUX4*-ERD (Fig 5.1A).

To determine the expression landscape induced by *DUX4* expression, a microarray of transcriptional changes induced by the *DUX4* constructs was performed (by Paul Knopp). Primary mouse satellite cells from three adult C57BL10 male mice were expanded, split and infected in parallel. RNA was prepared 20 hours later to assay early gene expression changes caused by *DUX4*, before more non-specific changes associated with cell death. Gene expression was analysed using Affymetrix GeneChip Mouse Gene 1.0 ST Arrays. Here we investigate this data to ascertain the genes and pathways perturbed by *DUX4* expression and to determine the impact of *DUX4* expression on signalling entropy.

5.2.1.2 DUX4 is a transcriptional activator The *DUX4* construct transcriptomic data was log-normalised using RMA, batch effects were detected by clustering analysis and corrected for using the ComBat function in R [378]. Hierarchical clustering and PCA on the 1000 most variable probes demonstrated that the different constructs clustered as expected. Full length *DUX4* clustered with the transcriptionally active tMAL*DUX4*-

VP16, while *DUX4c* and tMAL*DUX4* displayed similar transcriptional profiles, distinct from those obtained for *DUX4* and tMAL*DUX4*-VP16 (Fig 5.2). tMAL*DUX4*-ERD, designed to negatively regulate *DUX4* transcriptional targets, did not cluster with the other *DUX4* constructs.

We next performed a differential expression analysis. Using an empirical Bayes approach [269], we compared each *DUX4* construct to control retrovirus. By considering, t-values, corresponding to expression differences from control for each microarray probeset, we evaluated the concordance in transcription caused by each *DUX4* construct (Fig 5.3). We found that the *DUX4* and tMAL*DUX4*-VP16 constructs displayed a strong positive correlation in their transcriptional perturbations from control (Pearson's $r = 0.835$, $p < 2.2 \times 10^{-16}$), confirming that *DUX4* is a transcriptional activator. In contrast, *DUX4* and tMAL*DUX4*-ERD displayed no significant correlation in transcriptional perturbation from control (Pearson's $r = 0.007$, $p = 0.15$). The tMAL*DUX4*-VP16 and tMAL*DUX4*-ERD constructs did display significantly anti-correlated transcriptional perturbations from control, however, confirming the fact that the ERD and VP16 domains mediate inverse transcriptional responses. However, the anti-correlation between tMAL*DUX4*-VP16 and tMAL*DUX4*-ERD mediated transcription, though significant is weak in magnitude (Pearson's $r = -0.087$, $p < 2.2 \times 10^{-16}$). Coupled with the fact that the transcriptional changes induced by *DUX4* are not mirrored by the repressive tMAL*DUX4*-ERD, this suggests that *DUX4* is a transcriptional activator of genes not normally expressed in murine muscle satellite cells. Were this not the case one would expect the ERD construct to repress these genes and hence display a negative correlation with the *DUX4* construct.

The transcriptional changes caused by *DUX4c* were significantly positively correlated with those caused by *DUX4* (Pearson's $r = 0.41$, $p < 2.2 \times 10^{-16}$). Interestingly, we also found that *DUX4c* induced transcriptional changes which were significantly positively correlated with both tMAL*DUX4*-VP16 and tMAL*DUX4*-ERD (Pearson's $r = 0.55$, $p < 2.2 \times 10^{-16}$ and Pearson's $r = 0.04$, $p = 3.5 \times 10^{-12}$ respectively). This result suggests that whilst *DUX4c* transcriptionally activates some *DUX4* target genes, it may also repress others. This result is indicative of a possible antagonistic function for *DUX4c* on the *DUX4* phenotype. We also note that the transcriptional changes caused by tMAL*DUX4* and *DUX4c* are highly correlated (Pearson's $r = 0.817$, $p < 2.2 \times 10^{-16}$), as anticipated from the high sequence similarity of these two constructs and our clustering analysis.

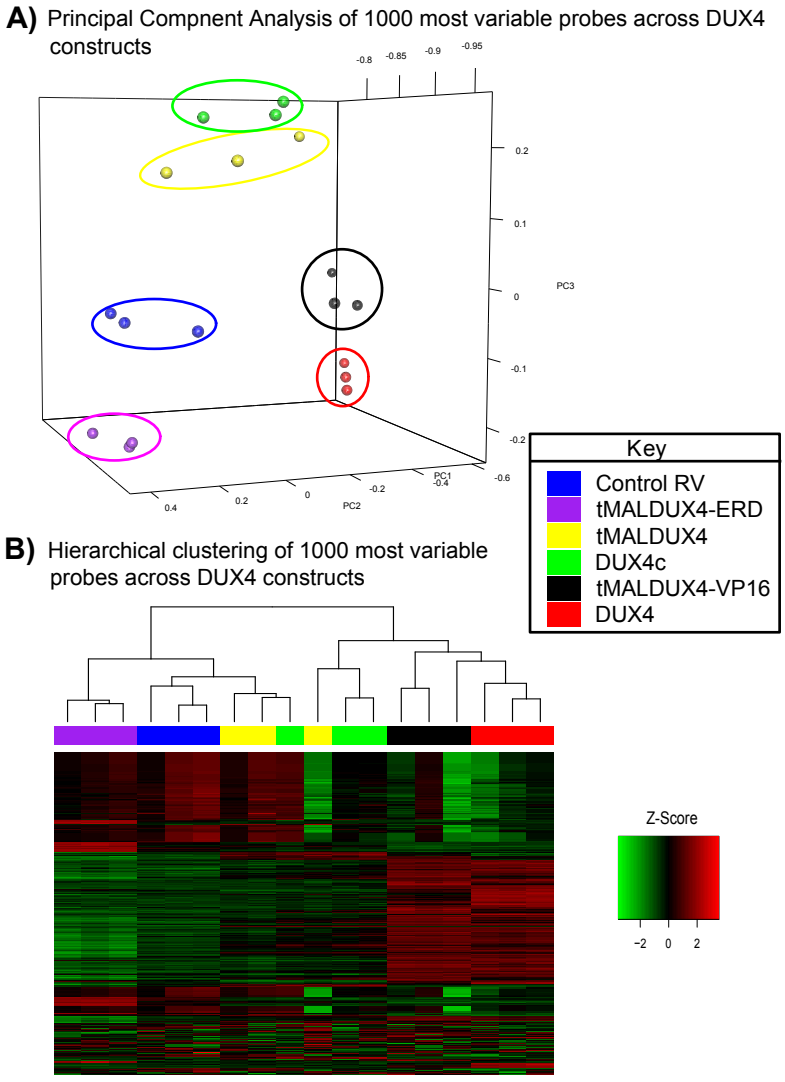

Figure 5.2: **Clustering of the *DUX4* construct infected myoblast samples.** (A) PCA and (B) hierarchical clustering on the 1000 most variable probes across the 5 *DUX4* retroviral constructs show the clustering of technical replicates, demonstrating reproducibility. Note the clustering of tMAL*DUX4*-VP16 infected samples with full length *DUX4* infected samples, and *DUX4c* infected samples with tMAL*DUX4* infected samples.

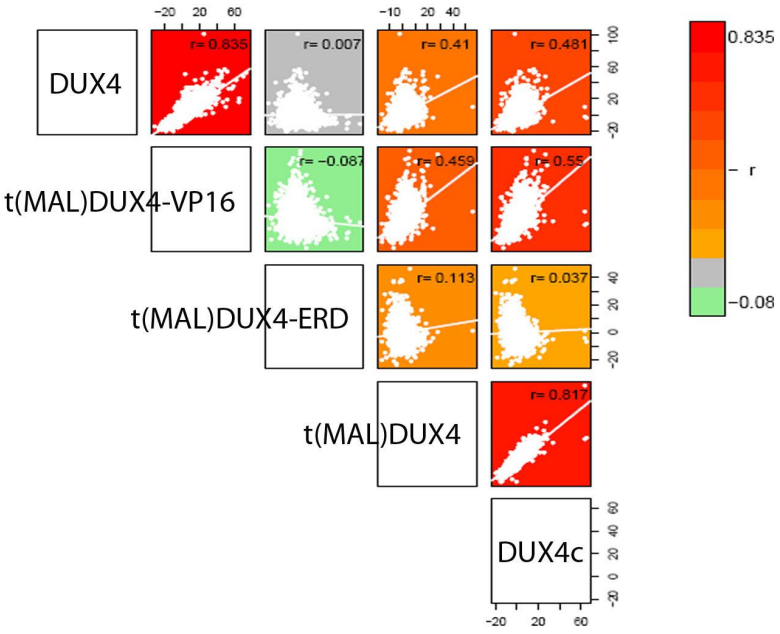

Figure 5.3: **Relationship between transcriptional perturbations induced by the *DUX4* constructs.** The plot shows correlations between differential expression t-values comparing each *DUX4* construct to control. Plots are coloured red to green according to correlation, grey plots did not display significant correlations. Of note is the negative correlation between tMAL*DUX4*-VP16 and tMAL*DUX4*-ERD and the positive correlations between *DUX4* and *DUX4c*.

5.2.1.3 *DUX4*, but not *DUX4c*, induces genes associated with apoptosis and reduced cell proliferation *DUX4* and *DUX4c* expression in murine satellite cells result in some similar phenotypes (inhibition of differentiation [218]) and some opposite phenotypes (apoptosis, proliferation [217]). Moreover, both the transcriptionally active tMAL*DUX4*-VP16 and the transcriptionally repressive tMAL*DUX4*-ERD result in transcriptomic profiles which are positively correlated with those induced by *DUX4c* expression, suggesting that *DUX4c*, acts both synergistically and antagonistically with *DUX4*. We thus investigated the pathways which are co-regulated and differentially regulated, by these two similar genes.

To identify these gene sets robustly we filtered genes sequentially, utilising information from all the 5 *DUX4* constructs. We employed a relaxed significance threshold for differential expression ($p < 0.05$) and considered for each of the 5 *DUX4* constructs compared to control independently. Taking advantage of construct similarity demonstrated above, a microarray probeset was considered significantly up-regulated by *DUX4* if it was up-regulated by both *DUX4* and tMAL*DUX4*-VP16 constructs. A probeset was considered significantly up-regulated by *DUX4c* if it was up-regulated by both *DUX4c* and tMAL*DUX4* (Fig 5.4 summarises the filtering process). After mapping probesets to unique RefSeq identifiers, we found 211 genes were co-up-regulated by *DUX4* and *DUX4c* whilst 245 genes were co-down-regulated.

We next performed a GSEA using a Fisher's exact test, to evaluate whether the genes commonly and differentially regulated by *DUX4* and *DUX4c* are significantly associated with particular functional classes (Materials and Methods, Chapter 5 [308, 368]). After correcting for multiple testing we found that gene sets down-regulated by both *DUX4* and *DUX4c* were significantly enriched for axonal guidance and muscle proteins, of interest, given that *DUX4* expression induces a neuronal phenotype in ES cells [226]. Conversely genes up-regulated by both *DUX4* and *DUX4c* were significantly enriched for urogenital, gland and vasculature development. The association with genital formation is intriguing given that *DUX4* is found to be expressed at high levels in the testes [216], and that FSHD has been found to show less severity in females than males [379]. Moreover, the association with blood vessel development, is of note given the association between FSHD and the presentation of a Coat's disease like retinal vasculopathy [186].

Genes down-regulated by *DUX4* but not *DUX4c* were significantly enriched for regulation of cell proliferation and apoptosis, in concordance with previous findings that *DUX4* but not *DUX4c* inhibits proliferation and increases cell death [217]. Moreover, considerable enrichment was found among these genes for nitrogen compound metabolic processes, indicating a role for nitric oxide in *DUX4* mediated oxidative stress sensitivity.

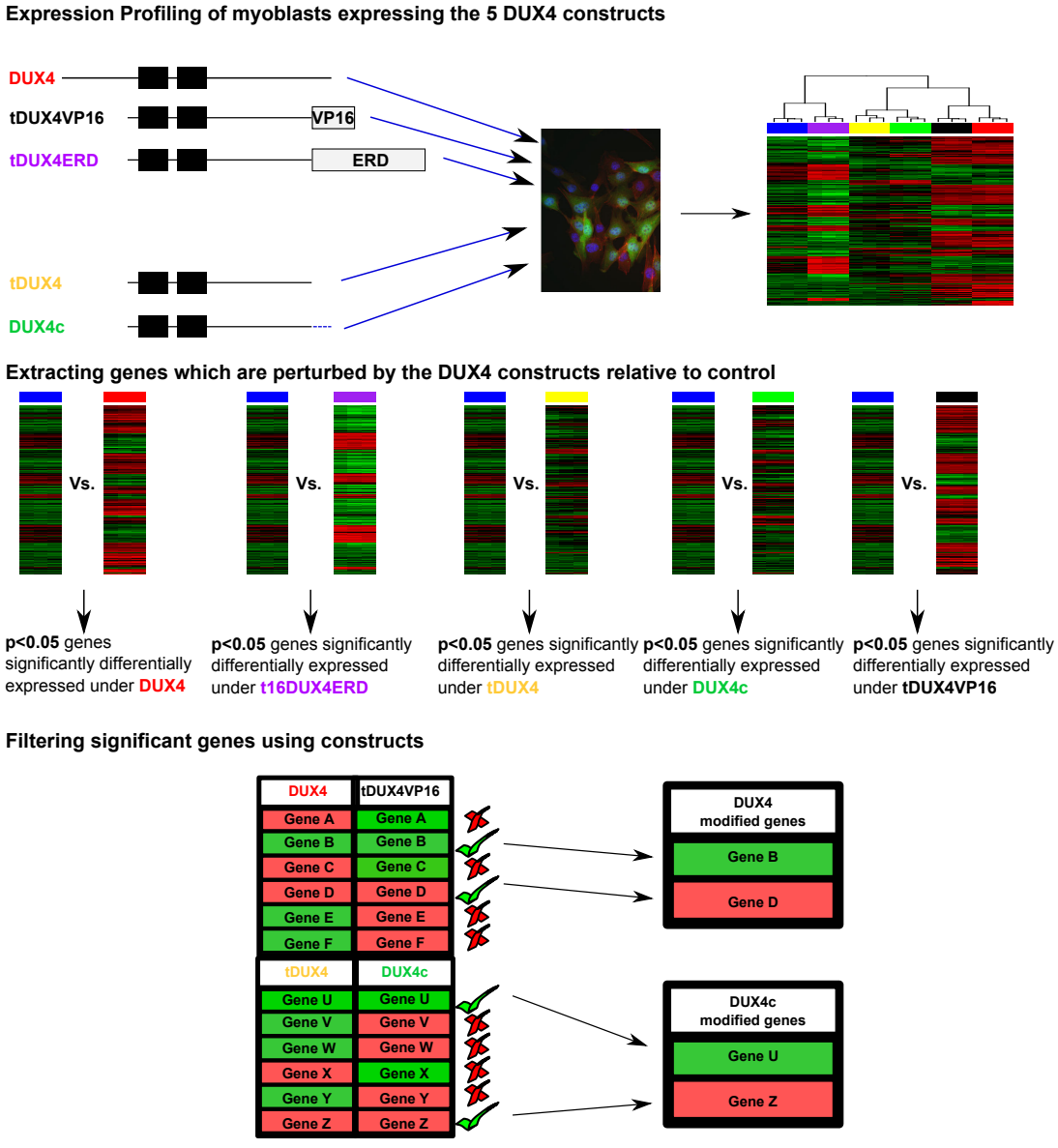

Figure 5.4: **Flowchart describing the selection of *DUX4* and *DUX4c* expressed genes.** A probeset is considered significantly up-regulated by *DUX4* if it is up-regulated by both *DUX4* and tMAL*DUX4*-VP16 constructs. A probeset is considered significantly up-regulated by *DUX4c* if it is up-regulated by both *DUX4c* and tMAL*DUX4*.

We note that genes up-regulated by *DUX4* and not *DUX4c* as well as genes regulated only by *DUX4c* showed no significantly enriched gene sets after correction for multiple testing.

5.2.1.4 A *DUX4* expression signature validates in multiple data sets Other groups have investigated the transcriptional perturbation induced by *DUX4* expression [235, 234]. However, the results of these investigations differ in regard to the main pathways perturbed and a comparison of published *DUX4* microarrays has not yet been performed. Moreover a robust marker of *DUX4* expression valid in different cell lines is currently lacking.
We derived a signature of *DUX4* expression from our multiple construct dataset. By considering genes regulated by *DUX4* (as described in Fig 5.4), we defined a sample specific *DUX4* expression score, as the statistic of the t-test evaluating whether the expression of the genes up-regulated by *DUX4* is higher than those down-regulated.
We considered our *DUX4* expression score on three further microarray datasets, one describing i*DUX4* C2C12 cells exposed to different levels of doxycyline, another describing *DUX4* over-expression in C2C12 cells by retrovial infection and yet another describing *DUX4* over-expression in human myoblasts by lentiviral infection. This revealed that the *DUX4* expression score was significantly elevated in samples over-expressing *DUX4* regardless of cell type and mode of over-expression (Fig 5.5), validating our score as a robust indicator of *DUX4* expression.

5.2.1.5 *DUX4* expression increases signalling entropy Given that *DUX4* expression has been shown to inhibit satellite cell differentiation [380], we next investigated whether the *DUX4* constructs induced a stem cell like transcriptomic profile. We previously demonstrated in Chapter 3 that signalling entropy (a measure of signalling pathway promiscuity, derived from the integration of gene expression data with a PIN) is a powerful measure of cell differentiation potential, valid across multiple lineages. Here we compute signalling entropy for each microarray sample describing gene expression induced by the 5 *DUX4* constructs and control retrovirus.
We found that expression profiles induced by *DUX4* and tMAL*DUX4*-VP16 displayed significantly higher signalling entropy than control retrovirus and *DUX4c* ($p < 0.005$ and $p < 0.0006$ respectively Fig 5.6), suggesting that *DUX4* expression results in a more stem-cell like transcriptomic profile. In contrast tMAL*DUX4*-ERD displayed a significantly lower signalling entropy than control retrovirus ($p < 0.04$), suggesting that repression of *DUX4* target genes results in a more differentiated expression regime. Interestingly tMAL*DUX4* and *DUX4c* displayed similar signalling entropy to control retrovirus, sug-

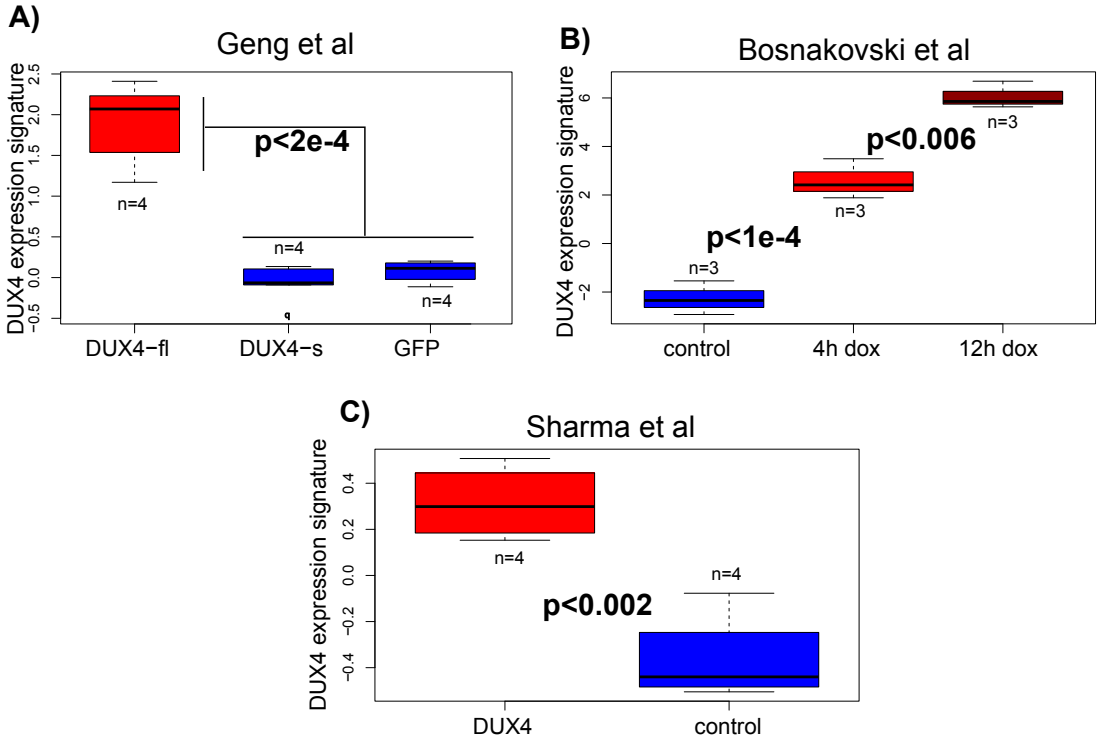

Figure 5.5: **A *DUX4* expression signature validates in multiple datasets.** We derived a signature of *DUX4* expression from our multiple construct data sets and demonstrated it capable of discriminating *DUX4* expressing samples from non-*DUX4* expressing samples in 3 independent micorarray datasets: (A) A data set produced by Geng *et al.*, of *DUX4* and a putative non-pathogenic splice variant *DUX4-s* over expression by lentivirus in human myoblasts. Control cells were infected with GFP. (B) A data set produced by Bosnakovski *et al.*, of i*DUX4* C2C12 murine myoblasts expressing *DUX4* in response to doxycycline exposure (4 hours (4h dox) and 12 hours (12h dox) exposure are shown, the longer exposure induces higher *DUX4*). (C) A data set produced by Sharma *et al.* of *DUX4* retroviral mediated over-expression in C2C12 murine myoblasts. p-values denote t-tests.

gesting that these constructs do not significantly alter global transcriptomic measures of differentiation potential.
We also computed signalling entropy in two independent *DUX4* data sets confirming that *DUX4* expression drives increased signalling entropy.

5.2.2 FSHD cell lines show an inhibition of differentiation

5.2.2.1 Overview In the previous section we considered over-expression of the primary FSHD candidate gene, *DUX4*, at high levels and the corresponding transcriptional perturbation, demonstrating via signalling entropy that *DUX4* induces a stem cell like phenotype. We also revealed that *DUX4* modifies the expression of genes involved in urogenital development, apoptosis and oxidative stress.
It has been shown in primary cultured myoblasts from FSHD patients that *DUX4* is expressed only at very low levels, though it is absent in myoblasts from healthy individuals [216]. Thus it is important to confirm our signalling entropy related finding that muscle differentiation is inhibited in FSHD patient derived cells. We could not demonstrate that signalling entropy is elevated in gene expression data corresponding to patient muscle biopsies, a finding which is likely due to the low expression of *DUX4* in these sample. Hence a more detailed experimental investigation is required to confirm that differentiation is indeed inhibited in FSHD.
FSHD is a highly heterogeneous pathology, however, and there are considerable differences in clinical presentation and *DUX4* expression levels, even between close relatives and monozygotic twins [376]. Hence it is important to establish models for FSHD which are as well controlled as possible. Recently such a model was derived by Krom *et al.* [241], in which a 54 year old male proband, mosaic for FSHD, displaying classical clinical presentation, underwent a biceps muscle biopsy. The biopsy contained two myoblast cell populations completely isogenic except for the D4Z4 repeat region, where one population contained a pathogenic truncation (3 D4Z4 repeats, characteristic of FSHD1) and the other did not (13 D4Z4 repeats). Myoblasts were isolated from the biopsy and immortalised via transduction with retrovirus encoding *hTERT* and *Cdk4*. Subsequently glass cylinders were utilised to isolate and expand single cells into clonal populations. These clonal populations were then genetically confirmed to contain either the pathological D4Z4 locus or the healthy.
These cell lines represent an important well controlled FSHD model, for the investigation of FSHD myogenesis. We thus obtained a healthy (54-6, 13 D4Z4 repeats) and a pathological (54-12, 3 D4Z4 repeats) cell line from this model. We first validate the lines as FSHD models, confirming that the pathological 54-12 cells over-express *DUX4*, and demonstrate

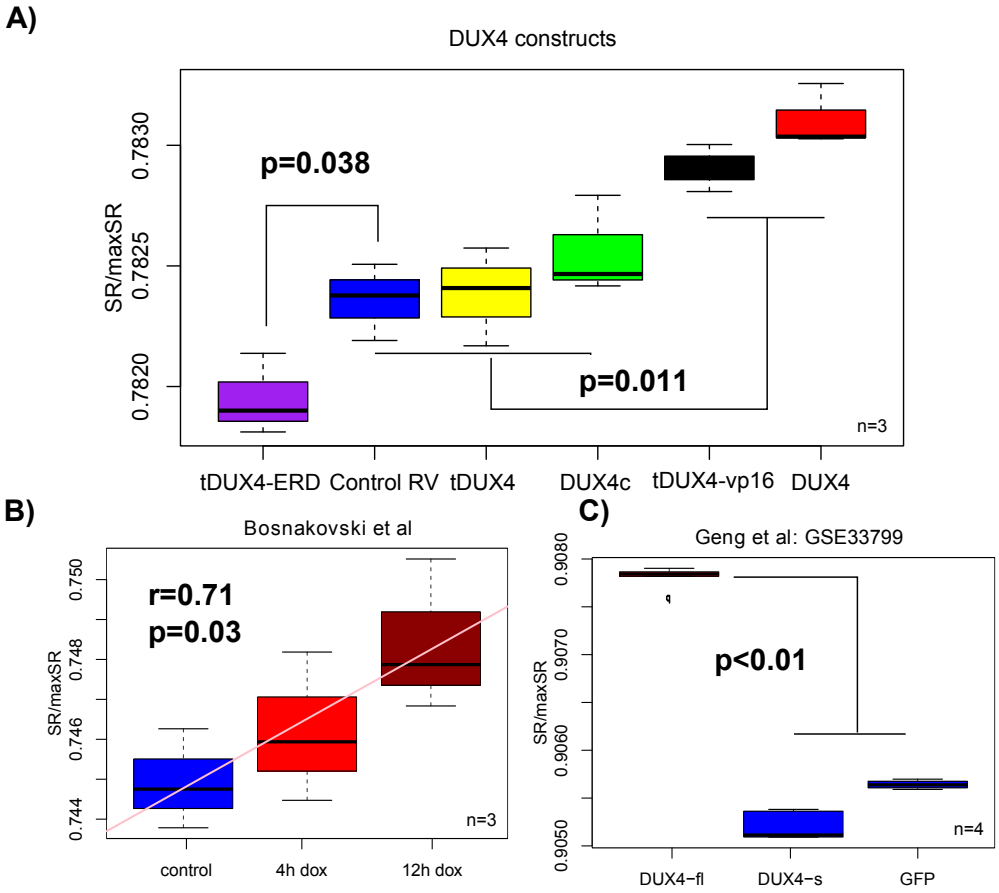

Figure 5.6: ***DUX4* expression increases signalling entropy.** (A) Signalling Entropy is elevated in the transcriptional profiles induced by *DUX4* and tMAL*DUX4*-VP16 expression, relative to control retrovirus, whilst it is reduced by tMAL*DUX4*-ERD expression. The p-values denote Wilcoxon tests (B) Signalling entropy is also correlated with exposure to *DUX4* inducing agent doxycycling in i*DUX4* cells, the p-value denotes linear regression. (C) Signalling entropy is elevated in *DUX4* expressing human myoblast samples. The p-value denotes a Wilcoxon test. These results support the hypothesis that *DUX4* inhibits myogenic differentiation.

defects in proliferation and myogenesis as well as a sensitivity to oxidative stress. Finally we examine the myogenesis defect in detail by high throughput time course imaging. We demonstrate that the 54-12 cells align and fuse more slowly than the healthy 54-6 cells, consistent with an inhibited differentiation driven by *DUX4* expression, as indicated by signalling entropy.

5.2.2.2 Pathological immortalised myoblast cell line 54-12 over expresses *DUX4* In the presenting paper, Krom *et al.* described *DUX4* expression in 5 healthy and 5 pathological lines isolated from the same mosaic patient, by qPCR. It was demonstrated that the average *DUX4* expression of the pathological lines was higher than the control, however, a direct comparison of the 54-6 control line and the 54-12 pathological line was not performed.
We demonstrated that 54-12 cells do express significantly more *DUX4* than 54-6 cells, although the expression level in 54-12 is extremely low (1.2×10^{-6} the level of the housekeeping gene *RPLPO*, Fig 5.7). It is worth noting that detection of *DUX4* by qPCR required the use of 4 times as much cDNA template than is normally used (50ng/20μl reaction), and cDNA needed to be prepared from high quality RNA (RIN score$>$ 9.5, $A_{260}/A_{280} = 1.9 - 2$). We were unable to detect any *DUX4* positive cells by immunocytochemistry.

5.2.2.3 FSHD cell lines demonstrate a proliferation defect Different studies have found contrasting results regarding FSHD myoblast proliferation, one study has reported no proliferation defect in FSHD myoblasts [235], whilst others have found that FSHD myoblasts undergo early cell cycle arrest [381] and that *DUX4* expression induces a proliferation defect [175]. To assess proliferation rate in our cell lines, cells were maintained in proliferation conditions and assessed for *5-Ethynyl-2'-deoxyuridine* (EdU) incorporation (Materials and Methods, Chapter 5). EdU is incorporated by myoblasts during S-phase of the cell cycle during the EdU incubation period (2 hours) and gives a measure of rate of proliferation.
There was a significantly reduced percentage of EdU nuclear uptake in the FSHD cell line 54-12 as compared to the control cell line 54-6 (Fig 5.8).

5.2.2.4 FSHD cell line 54-12 demonstrates a sensitivity to oxidative stress Sensitivity of proliferating FSHD myoblasts to oxidative stress is a well-documented phenotype [381, 175]. To assess the sensitivity of our cell lines to oxidative stress we incubated proliferating cells, plated at a density of 5×10^3 per well in a 96 well plate, with and with-

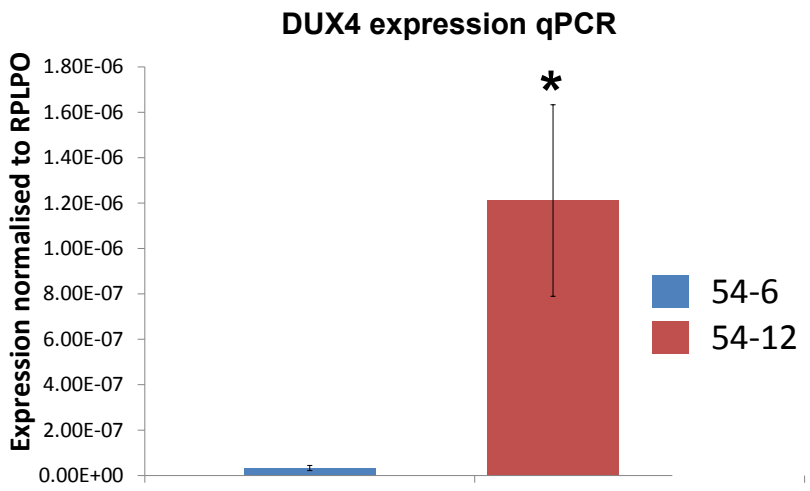

Figure 5.7: ***DUX4* is over expressed in pathological cell line 54-12.** qPCR evaluating *DUX4* expression in 54-12 and 54-6 immortalised myoblast cell lines, expression was normalised to housekeeping gene *RPLPO*. The assay was performed in triplicate, error bars denote standard error, significance was assessed via a two tailed, unpaired t-test, * denotes $p < 0.05$.

out 400μM hydrogen peroxide (H_2O_2) for 24 hours. Similar conditions have previously been utilised to induce oxidative stress in i*DUX4* myoblasts. Immunocytochemistry was performed to quantify morphological changes (via tubulin expression) and cell number alterations (*4',6-diamidino-2-phenylindole* (DAPI) counts) brought about by oxidative stress (Materials and Methods, Chapter 5).

Fold change in cell number between control and stressed conditions were computed for each cell line and compared. FSHD line 54-12 showed a significantly greater reduction in cell number under oxidative stress than the control line 54-6, demonstrating a clear oxidative stress sensitivity phenotype (Fig 5.9A).

Cell morphology was analysed using image analysis software written in R, using the EBImage package for automated cell phenotyping [382]. In particular the average eccentricity of cells (a measure of how stretched as opposed to circular cells are) was quantified in both stressed and unstressed conditions (Materials and Methods, Chapter 5).

The 54-12 cell line also showed a significantly greater increase in eccentricity under oxidative stress than observed in control (Fig 5.9B).

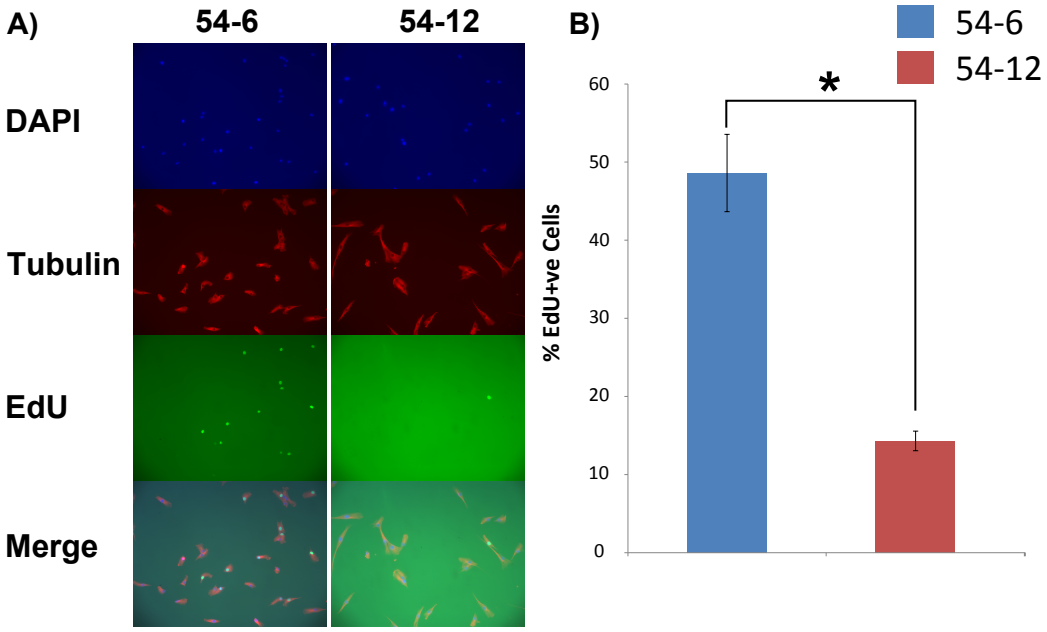

Figure 5.8: **Pathological cell line 54-12 displays a defect in proliferation.** (A) Immunocytochemistry: Cells were stained with DAPI, which binds to DNA, (blue), tubulin (red) and with a flurophore which binds to EdU (green). Note the elongated morphology of the pathological cells and the low nuclear EdU incorporation. B) Quantification of nuclear EdU incorporation reveals a defect in proliferation in the 54-12 cells. The assay was performed in triplicate, at least 500 cells were imaged in 5 distinct fields per repeat, error bars denote standard error, significance was assessed via a two tailed, unpaired t-test, * denotes $p < 0.05$.

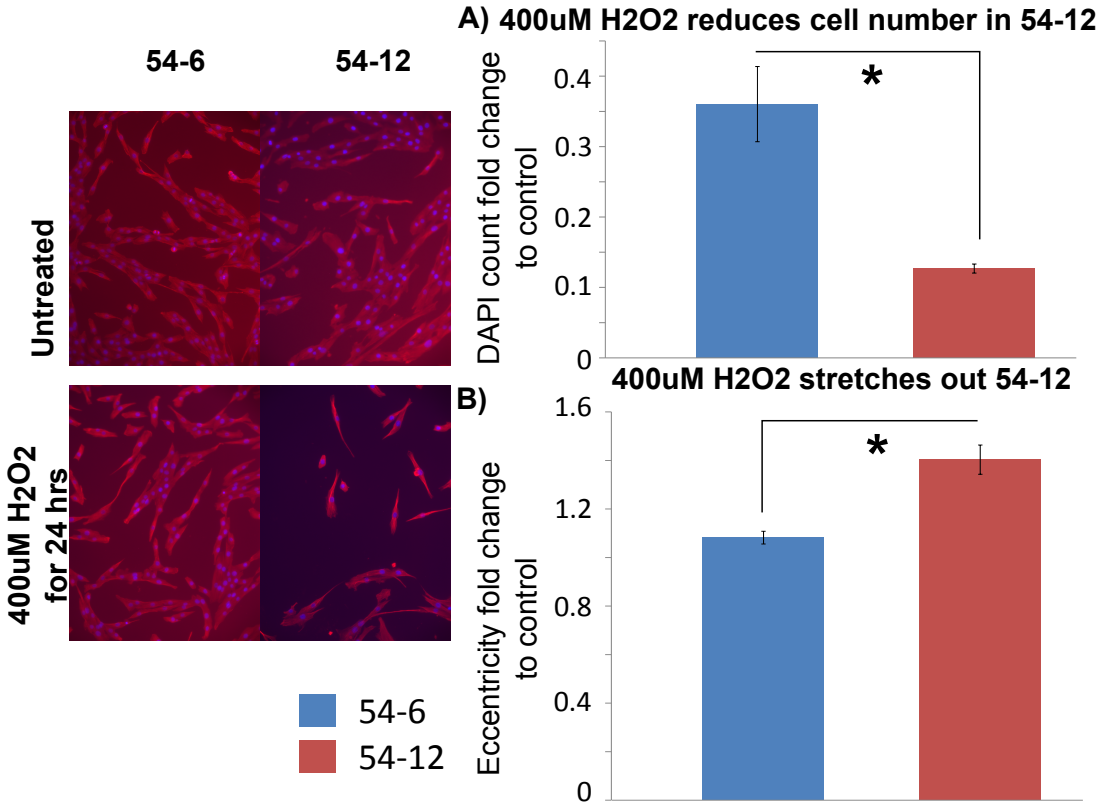

Figure 5.9: **Pathological cell line 54-12 displays a sensitivity to oxidative stress.** Immunocytochemistry: Cells were stained with DAPI, which binds to DNA, (blue) and tubulin (red) to elucidate morphology. (A) 54-12 cells show a significantly greater reduction in cell number following oxidative stress as compared to the healthy cell line 54-6. (B) 54-12 cell lines undergo a significantly greater increase in cell eccentricity following oxidative stress than the healthy cell line. The assay was performed in triplicate, at least 500 cells were imaged in 5 distinct fields per repeat, error bars denote standard error, significance was assessed via a two tailed, unpaired t-test, * denotes $p < 0.05$.

5.2.2.5 FSHD cell line 54-12 displays an atrophic myotube phenotype Myotube morphology has been reported as considerably heterogeneous in FSHD, with some patient derived myoblasts displaying a disorganised myotube phenotype and others an atrophic phenotype [235, 236, 375]. Moreover, the formation of atrophic myotubes has been linked directly to *DUX4* expression [375]. To assess the myotube phenotype in our cell lines in a robust dynamic manner, we placed confluent human myoblasts into mitogen poor media to induce differentiation and fusion and fixed the cells after 2, 3 and 5 days. Immunocytochemistry was then performed to evaluate the expression of myosin heavy chain (*MyHC*) using the MF-20 antibody (Materials and Methods, Chapter 5).
Image analysis software was written using the EBImage package in R to evaluate the fusion index for each cell line as well as the area of image positive for *MyHC*, to assess myotube morphology (Materials and Methods, Chapter 5).
In line with the results of Krom *et al.*, fusion index appeared to be higher in the 54-12 cell line as compared to the 54-6. Quantification of myotube size by *MyHC* positive area, further revealed that the 54-12 cell line, displayed significantly smaller myotubes than 54-6 at all time points, indicative of an atrophic phenotype (Fig 5.10).

5.2.2.6 Immortalised cell lines 54-6 and 54-12 fuse at different rates To investigate in detail the differences between the myogenesis programmes in the pathological and healthy cell lines, we performed high density, time lapse microscopy imaging, in triplicate, over 5 days of differentiation. One 10× magnification image of the cells was captured every 5 minutes over the process using an an Eclipse Ti-E Live Cell Imaging System, generating a total of 8640 images (Materials and Methods, Chapter 5).
We subsequently developed image analysis software to process and analyse the images, generating a time course of morphological changes (eccentricity/elongation of cells) in each cell line over the myogenesis programme (Materials and Methods, Chapter 5).
The time course imaging revealed the myogenesis process in great detail, we found that cells underwent a rapid alignment phase (increasing eccentricity), followed by a cytoplasmic expansion during fusion, during which unfused plated cells were pushed off the plate, causing them to round up (decreasing eccentricity). Following this rapid phase of cytoplasmic expansion, the first visible myotubes appeared (increasing eccentricity, Fig 5.11). Evaluation of differences between eccentricities at each time point using an empirical Bayes approach [269] revealed significant morphological differences between cell lines occurred particularly during the early stages of myogenesis. We found that both the alignment and cytoplasmic expansion phases took slightly longer and resulted in less extreme morphological changes in the pathological line, 54-12, as compared to the healthy line,

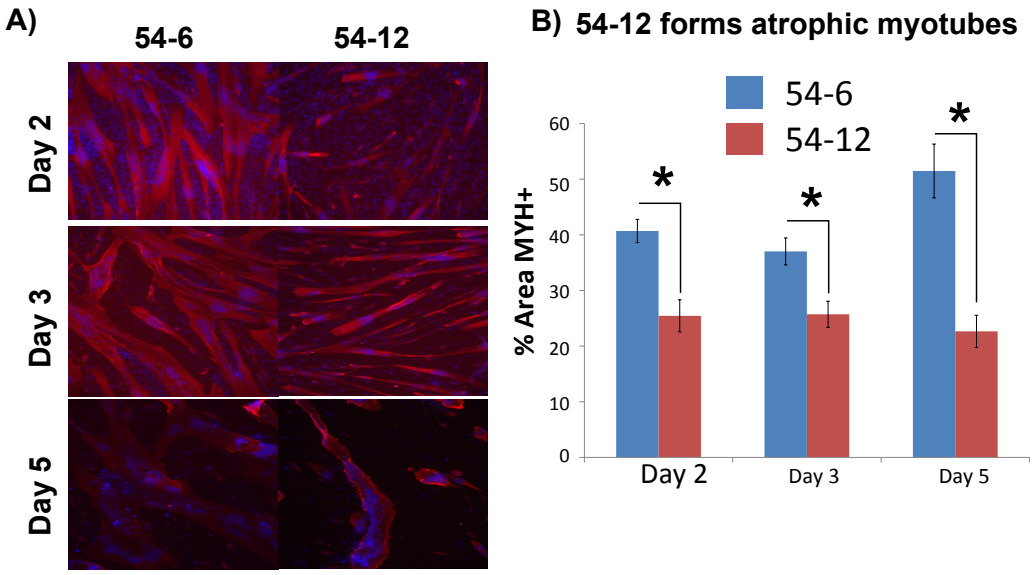

Figure 5.10: **Pathological cell line 54-12 displays an atrophic myotube phenotype.** (A) Cells were stained with DAPI, which binds to DNA, (blue) and *MyHC* (red) to elucidate morphology. (B) 54-12 cell lines differentiate into significantly smaller myotubes than the healthy line 54-6. The assay was performed in triplicate, at least 500 cells were imaged in 5 distinct fields per repeat, error bars denote standard error, significance was assessed via a two tailed, unpaired t-test, * denotes $p < 0.05$.

54-6.
This retardation of myogenic progression in FSHD is coherent with an inhibition of differentiation caused by *DUX4* expression.

5.2.3 Network Rewiring in FSHD reveals β-catenin as central to *DUX4* driven pathomechanisms

5.2.3.1 Overview We have shown that *DUX4* expression induces a high signalling entropy and that FSHD myoblasts have an impaired myogenesis programme. However, *DUX4* is only expressed at very low levels in FSHD myoblasts. A hypothesis surrounding this is that transient high *DUX4* expression, early in development, may pre-programme FSHD myoblasts to undergo aberrant differentiation in a hysteretic manner, before being epigenetically silenced. If this is the case the low *DUX4* expression in FSHD myoblasts may represent inefficient silencing. Such a hypothesis raises the question as to whether *DUX4* is the optimal target for therapeutic intervention, and motivates the investigation of more downstream targets that may be responsible for perpetuating the adverse effects of *DUX4* expression in FSHD muscle.
In this section we consider gene expression data corresponding to FSHD primary muscle biopsies from four independent data sets. To elucidate the drivers of FSHD pathomechansisms we employ a local network theoretic tool introduced in Chapter 2: InSpiRe.
The InSpiRe algorithm employs a local entropy measure of network rewiring as well as a symmetrised Kullback-Leibler divergence alongside statistical graph sparsification to extract a subset of the human interactome with is rewired between two phenotypes described by expression data.
We apply InSpiRe to a meta-analysis of FSHD and related gene expression data sets, to extract a network describing protein interactions which are specifically perturbed in FSHD muscle and which are not attributable to muscle atrophy, inflammation, ageing or other muscle diseases. Importantly, we demonstrate that the genes in our FSHD network are significantly perturbed by *DUX4* expression. Finally we utilise betweenness centrality to identify β-catenin as a critical bottleneck to FSHD pathological signalling and demonstrate experimentally that downstream targets of β-catenin signalling are perturbed by *DUX4* expression.

5.2.3.2 Meta-analysis of FSHD data sets using InSpiRe InSpiRe is a differential network methodology we designed to extract a subset of the human PIN containing proteins and interactions that are altered between two phenotypes described by expression data and is explained in detail in Chapter 2.

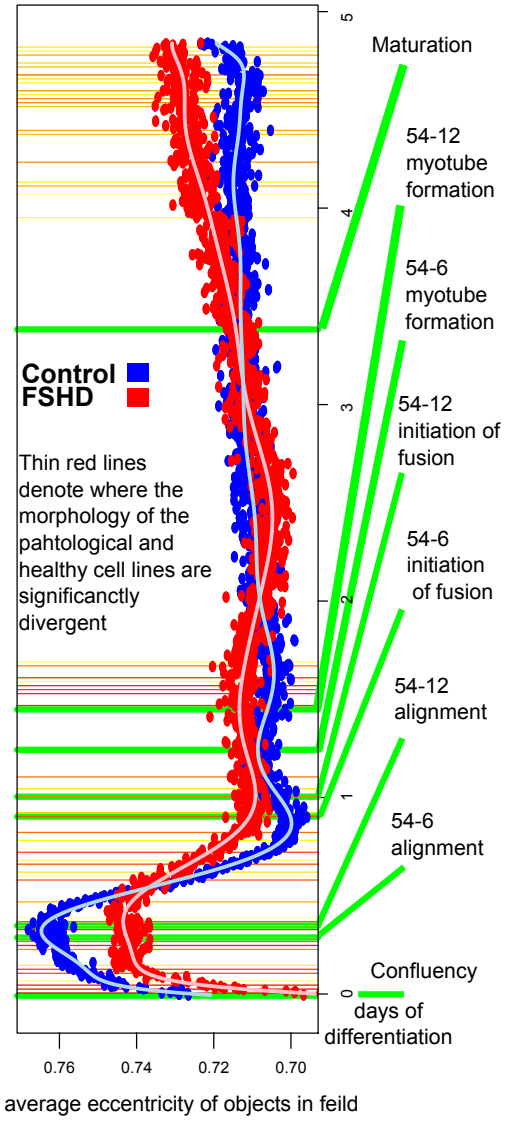

Figure 5.11: **Pathological cell line 54-12 aligns and fuses more slowly than control line 54-6.** A plot of average cell eccentricity (elongation) over the myoblast differentiation time course. Thin lines denote time points where the FSHD and control cell lines significantly differ morphologically ($p < 0.05$) and are coloured red to yellow in order of significance. Green lines denote morphologically important time points during the myogenesis programme.

We ran the first two steps of InSpiRe on each of 4 FSHD data sets obtained from the GEO Database [270] (GSE3307 and GSE10760 plus GSE26145 combined with GSE26061 but divided into myoblasts and myotubes. Materials and Methods, Chapter 5). ~ 3500 genes were implicated per set by InSpiRe as rewiring between FSHD and control samples. Considering all FSHD data, the intersection of rewired sets consisted of a significantly large overlap of 829 genes ($p < 10^{-5}$, based on randomly selecting genes from each data set and assessing the size of overlap obtained). Many genes in this intersection have been implicated in FSHD, *e.g. TP53* [181], *JUNB* [52, 213], *HIF1A* [52], *WNT3* [186], *LMO3* [52], *ANXA4* [187] and *HSPB1* [383]. GSEA [368] on the intersection implicated many enriched pathways associated with FSHD, including myogenesis [213], p53 signalling [181], Wnt pathway [186], regulation of actin cytoskeleton [377] and *VEGF* signalling pathway [187].

5.2.3.3 Gene expression changes specific to FSHD Superimposed on network rewiring due to FSHD molecular mechanisms, are rewiring events due to non-specific changes. To identify genes implicated as rewiring specifically in FSHD, we also ran the first two steps of InSpiRe on two data sets each describing skeletal muscle gene expression during ageing (GSE5086 and GSE9676), disuse atrophy (GSE5110 and GSE8872) and other muscle diseases involving inflammation and wasting (GSE3307, where Juvenile dermatomyositis and limb-girdle muscular dystrophy type 2A data sets were independently analysed). Genes rewiring in these non-FSHD data sets were considered as secondary rewiring. Of the 829 InSpiRe identified genes, 273 were associated with ageing, 364 with disuse atrophy and 394 with other muscle diseases, identifying a set of 164 genes specifically rewiring in FSHD. For the final stage of InSpiRe we considered the 164 high confidence genes and their direct neighbours in the protein interaction network (a complete network of 2866 genes). Statistical sparsification on the largest FSHD data set (GSE10760) was performed to eliminate interactions that were not significantly altered between FSHD and control phenotypes to generate the FSHD network.

5.2.3.4 The FSHD network Our FSHD network consists of 2616 proteins and 15972 interactions, the majority of which forms a single maximally connected component (2603/2616 genes). To evaluate the significance of certain properties of the FSHD network, we computed a distribution of random networks by performing 1000 random selections of 164 genes and re-running the statistical sparsification on GSE10760 each time. This demonstrated that the FSHD network has significantly more interactions and genes than one would expect by chance ($p = 0.04$ and $p = 0.034$ respectively). Such network density implies that signalling dysregulation underlying FSHD is a coordinated

perturbation of a large number of intersecting signalling pathways.
The FSHD network is provided in a Cytoscape format as supplementary material to our publication [384].

5.2.3.5 *DUX4*-driven gene expression mirrors FSHD We next determined how much network rewiring in FSHD is directly due to *DUX4*, to investigate this we re-considered the microarray dataset, describing transcriptional perturbations induced by a panel of *DUX4* retroviral constructs. To establish whether genes in the FSHD network are perturbed as a result of *DUX4* expression, we extracted all microarray probes mapping to genes with direct human orthologs in our FSHD network, creating a network probeset consisting of 1866 genes. We then performed a re-sampling procedure (see Materials and Methods, Chapter 5) to determine whether the expression of the network probeset was a significant biomarker of *DUX4* expression. This analysis confirmed that mouse orthologs of genes in our human FSHD network are significantly modified by *DUX4* ($p < 10^{-5}$).

5.2.3.6 Dysregulation of β-catenin signalling is central to rewiring in FSHD To identify critical genes and pathways in our FSHD network we employed local network measures. Betweenness centrality measures the number of shortest paths between any two genes passing through a given gene, and can identify signalling bottlenecks. Genes in our network demonstrating high betweenness centrality are important for coordination of signal dysregulation in FSHD: the gene with the highest is *CTNNB1* encoding β-catenin. *CTNNB1* is also highly connected in our network, with a degree of 73, supporting a role for this gene in numerous dysregulated interactions. To determine if an increase in β-catenin activity is occurring in FSHD muscle we considered the neighbourhood of *CTNNB1* in the FSHD network (Fig 5.12). β-catenin is highly correlated with its neighbours in the FSHD network across FSHD samples, but not across control samples, implying an increase in β-catenin activation in FSHD (Fig 5.12).
To determine whether β-catenin is acting via its role in transcription, we queried our network for downstream targets that mediate β-catenin signalling, *i.e.* the TCF/LEF family of transcription factors. Importantly, all members of the TCF/LEF family (*TCF7*, *TCF7L1* (TCF-3), *TCF7L2* (TCF-4) and *LEF1*) as well as *PITX2* were involved in network rewiring in FSHD.
To analyse control of β-catenin activity via canonical Wnt signalling, we queried our network for Wnt, dishevelled and frizzled family members [385]. This revealed *WNT16*, two dishevelled *DVL1* and *DVL2* and one frizzled *FZD1*.
These genes upstream and downstream of β-catenin are connected in the FSHD network (except *WNT16*), indicating dysregulation of the β-catenin signalling pathway is con-

tributing to FSHD pathogenesis. There is a significantly increased positive correlation in gene expression along the chain: *FZD1* $\rightarrow$ *DVL1* $\rightarrow$ *CTNNB1* $\rightarrow$ TCF-3 (*TCFL1*) $\rightarrow$ c-Myc (*MYC*) in FSHD samples, implying an increased activation of this pathway (Fig 5.13). Increased negative correlation between β-catenin gene expression and that of *PITX2* and increased positive correlation between gene expression of *PITX2* and *LEF1* also occurred. β-catenin is also involved in numerous other processes including substantially altered correlations with *CASP3*, *CASP8*, interactions associated with apoptosis, and Hypoxia inducible factor 1-α (HIF1-α) (Fig 5.13).

5.2.3.7 HIF1-α Signalling HIF1-α is one of the most rewired of the 164 genes and increases in activity in FSHD (Fig 5.14), with many genes associated with HIF1-α signalling in the FSHD network, including, *VHL, HSP90AA1, RBX1, RRAS, VEGFA, MAPK8, NCOA1, PIK3R3, SLC2A4, HIF1AN* and *TCEB2*. HIF1-α signalling has recently been implicated as FSHD associated, due to the identification of several downstream components of the pathway as differentially expressed [52].

5.2.3.8 TNF-α over-activation of reactive oxygen species induced JNK cell death pathways Many genes involved in TNF-α over-activation of *Reactive Oxygen Species* (ROS) induced JNK cell death pathways were in our FSHD network. These included *MAP4K5, PARP2, JUNB* (Fig 5.15-5.17) *TNFA, JUN, JUND, JNK1* and *JNK3*. *MAP4K5* is a highly specific activator of JNK signalling [386], and displays a significantly increased ($p = 0.0035$) negative expression correlation with TNF-receptor associated factor 2 (*TRAF2*) in FSHD samples. *PARP2* is necessary for activating TNF-α induced necrosis [387] and displays increased positive expression correlation with *BRCA1* across FSHD samples, in an interaction associated with cell death [154]. *JUNB* displays significantly increased positive expression correlation with *FOS* ($p = 1.5 \times 10^8$) and *JUN* ($p = 0.044$). *JNK1* and *JNK3* display clear shifts from predominantly uncorrelated in expression with neighbours in control samples, to highly correlated in FSHD samples, implying their increased activity in FSHD muscle.

5.2.3.9 Perturbed Wnt/β-catenin signalling in *DUX4*-infected satellite cell-derived myoblasts Several Wnt/β-catenin targets were identified as being altered in both the human FSHD network and the *DUX4* microarray. These included Leucine rich repeat containing G protein coupled receptors 5 and 6 (*Lgr5/6*), Transcription factors 3 and 4 (*Tcf3/4*), Myogenic factor 5 (*Myf5* and Lymphoid enhancer binding factor 1 (*Lef1*). To validate this we performed qPCR on satellite cell-derived myoblasts retrovirally infected with the *DUX4* and control retroviral constructs for 24 and 48 hours. *Lef1,*

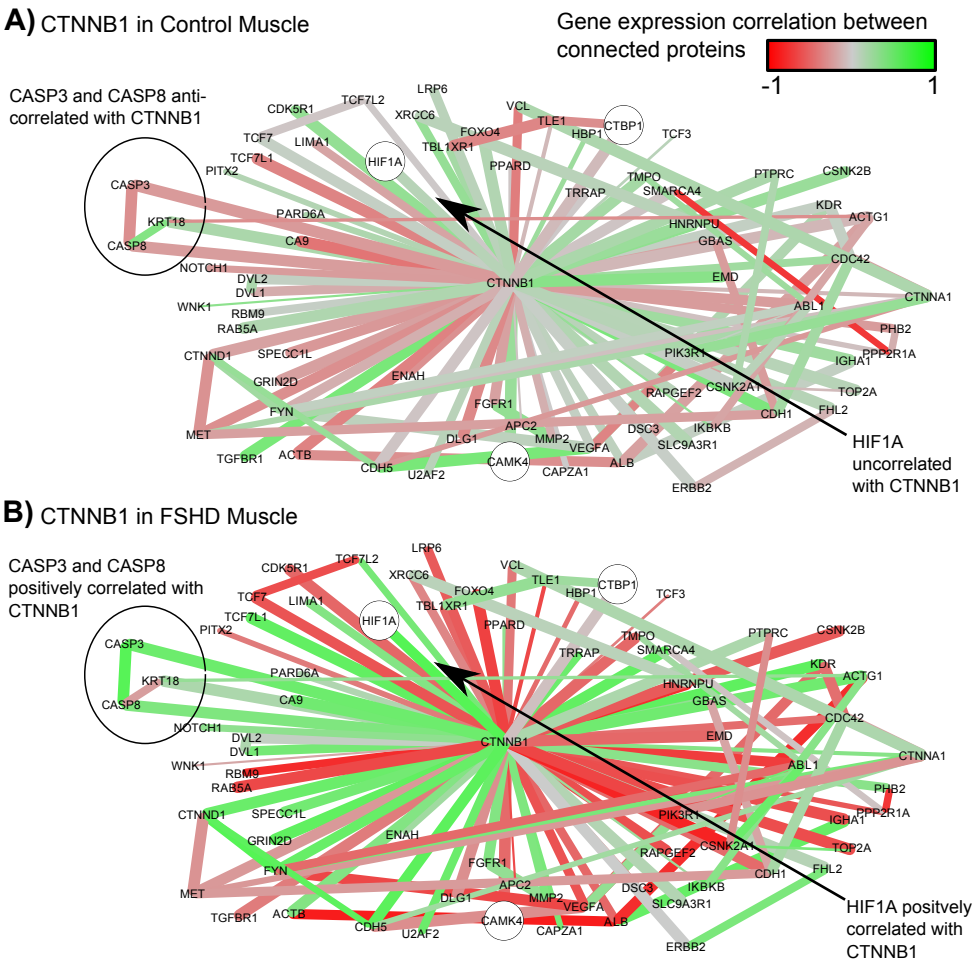

Figure 5.12: **The neighbourhood of *CTNNB1* in the FSHD network.** Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Red edges are negatively correlated, grey edges uncorrelated and green edges positively correlated. The thickness of edges is proportional to $1 - p$, where $p \in (0, 0.05]$ is the p-value of the statistical analysis performed to determine whether the correlation in gene expression between connected edges is different between FSHD and control samples. Large nodes belong to the set of 164 high confidence FSHD specific rewiring genes. There is a clear shift from predominantly uncorrelated to highly correlated between FSHD and controls, with an increased correlation between *CTNNB1* and its interaction partners across FSHD samples as compared to controls. This is indicative of increased β-catenin activity. Note the increased positive correlation between *CTNNB1* and *HIF1A*, *CASP3* and *CASP8*.

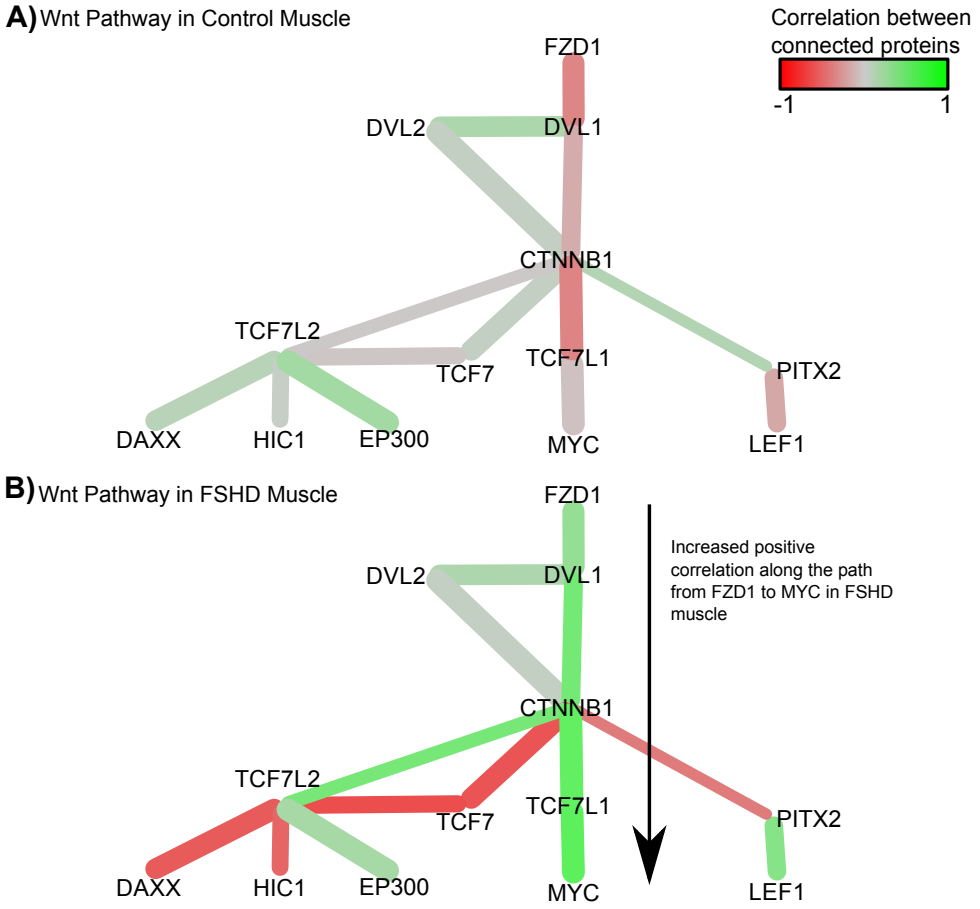

Figure 5.13: **The Wnt pathway in the FSHD network.** Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Note the shift from predominantly uncorrelated links in controls to highly correlated links in FSHD, implying an activation of this Wnt pathway in FSHD.

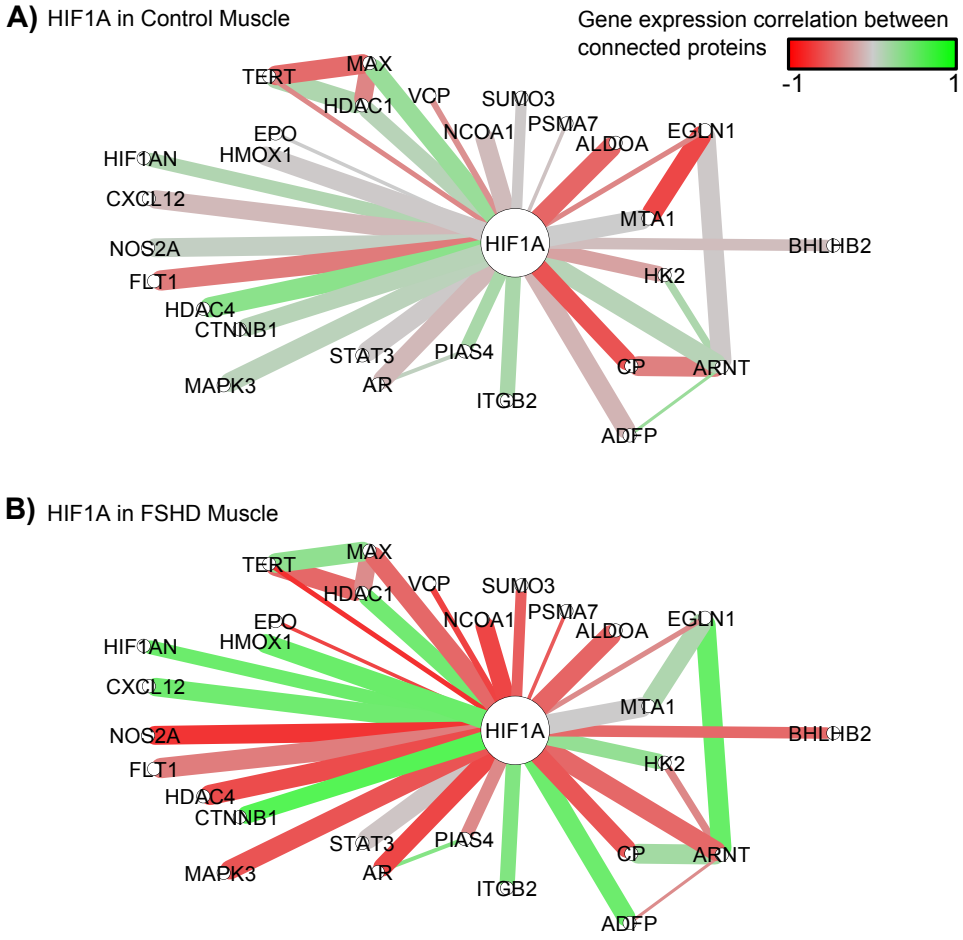

Figure 5.14: **The neighbourhood of *HIF1A* in the FSHD network.** Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Note the strong increase in correlation between *HIF1A* and *CTNNB1*.

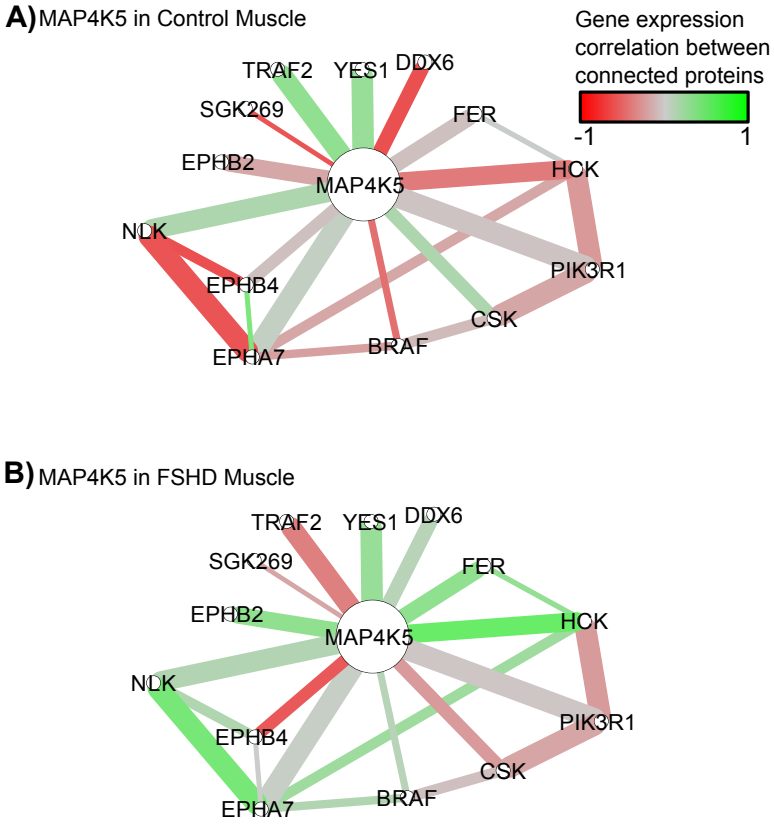

Figure 5.15: **The neighbourhood of *MAP4K5* in the FSHD network.** Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Note the strong increase in correlation between *MAP4K5* and *TRAF2*.

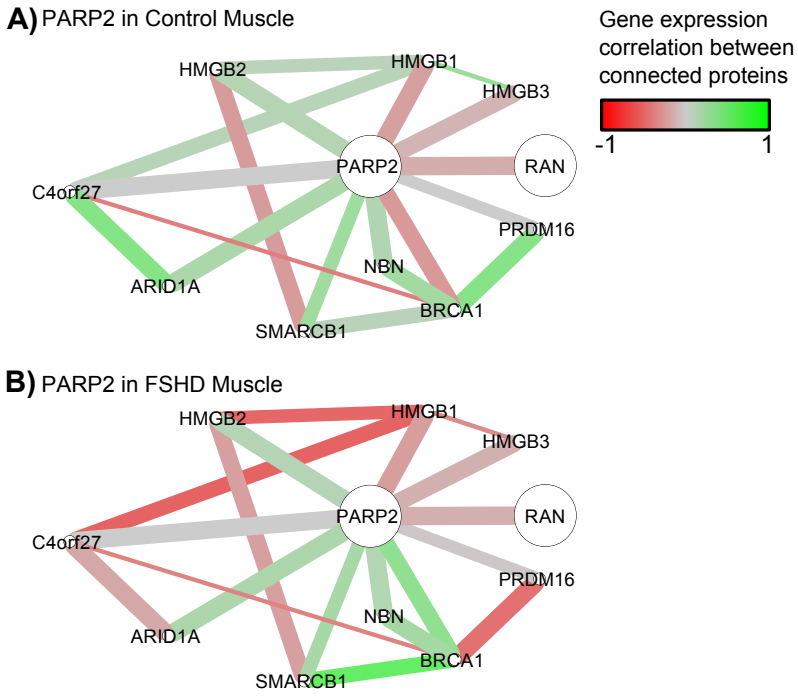

Figure 5.16: **The neighbourhood of *PARP2* in the FSHD network.** Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). Note the altered interaction between *PARP2* and *BRCA1* in FSHD.

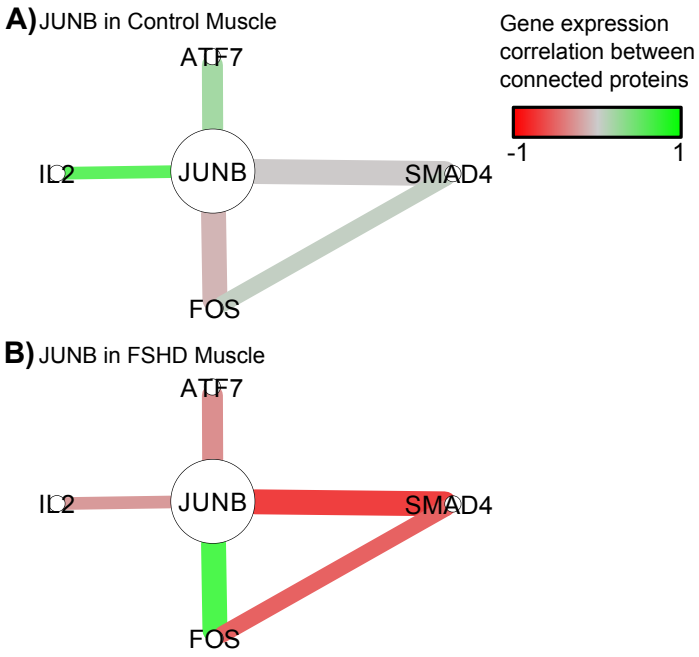

Figure 5.17: **The neighbourhood of *JUNB* in the FSHD network.** Interactions are coloured proportional to the Pearson correlation in gene expression between connected genes across control samples (A) and FSHD samples (B). The interaction between *JUNB* and *FOS* is significantly altered in FSHD muscle.

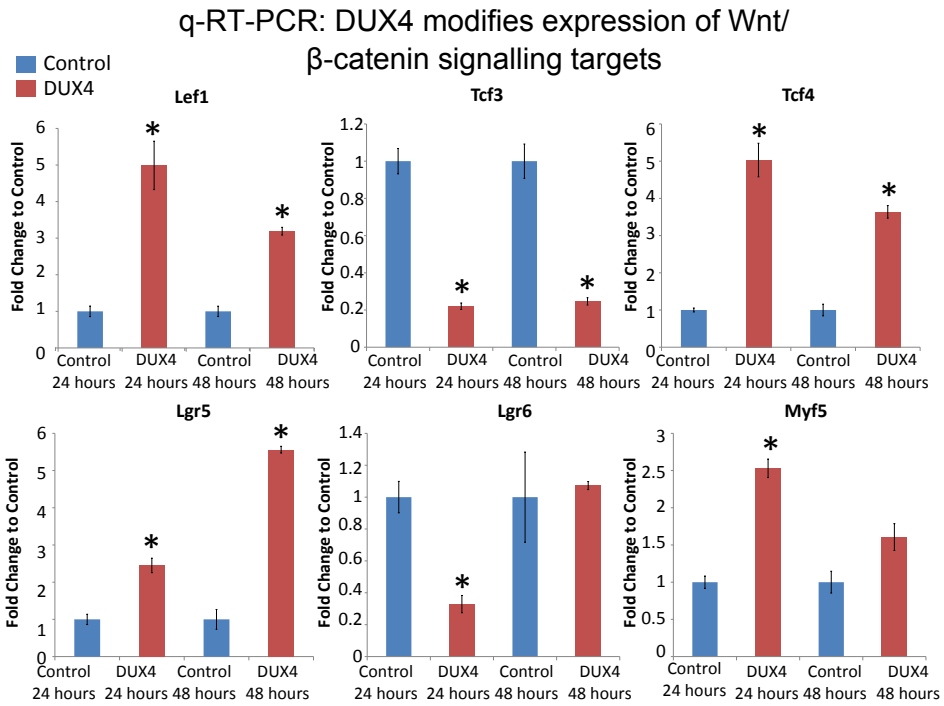

Figure 5.18: ***DUX4* perturbs Wnt/β-catenin signalling.** qPCR of downstream targets of Wnt/*beta*-catenin signalling *Lef1*, *Tcf3*, *Tcf4*, *Lgr5*, *Lgr6* and *Myf5* in satellite cell derived myoblasts, 24 and 48 hours after infection with *DUX4* and control retroviral constructs. Expression is displayed relative to levels in myoblasts infected with control retrovirus; $^*p < 0.05$.

Tcf4, *Lgr5* and *Myf5* were all significantly increased in *DUX4*-infected samples, whereas *Tcf3* and *Lgr6* were significantly decreased by *DUX4* (Fig 5.18), confirming that *DUX4* alters this signalling cascade at a transcriptional level.

5.2.3.10 Comparison of InSpiRe to other methodologies We compared InSpiRe to other commonly used network methodologies NetWalk [16] and GSEA on differentially expressed genes (described in detail in Materials and Methods, Chapter 5). NetWalk also uses a PIN to identify candidate genes via weighted random walks. Both NetWalk and InSpiRe displayed a significant consensus in the features identified across the four FSHD data sets ($p < 10^{-5}$), whereas no differentially expressed genes were consistently identified. This demonstrates the power of network based methodologies over conventional differential expression. NetWalk performed similarly to InSpiRe in identifying genes which

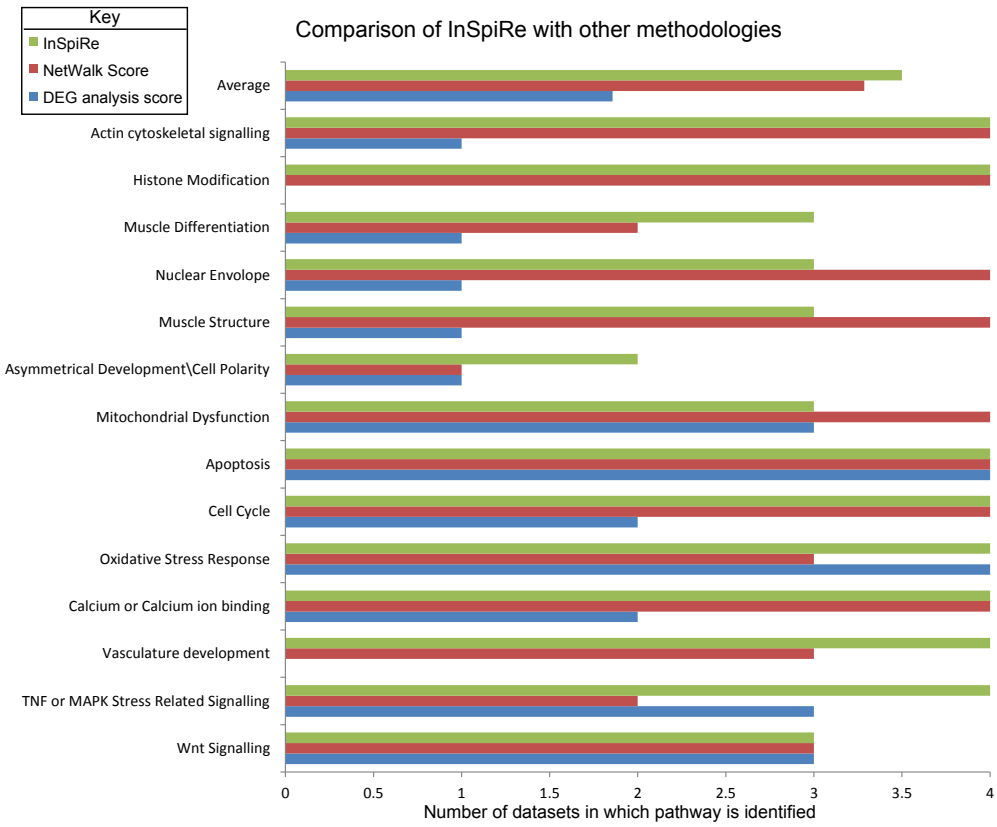

Figure 5.19: **Comparing InSpiRe to NetWalk and GSEA on differentially expressed genes.** Each method is scored with the number of data sets considered in which it identifies a given pathway. The average score across all pathways is highest for InSpiRe.

proved a significant classifier of the *DUX4* construct microarray samples ($p < 10^{-5}$) and generated a network with more genes and fewer connections than InSpiRe (3363 genes and 5651 connections). However, both NetWalk and GSEA on differentially expressed genes were inferior to InSpiRe in identifying functional annotations previously associated with FSHD (Fig 5.19).

5.3 Discussion

The aim of this chapter was to understand FSHD pathomechanisms, through the use of network theoretic tools and experimentation, in a manner related to *DUX4* expression.

We wished to understand the nature of the FSHD myogenesis phenotype through signalling entropy and to posit novel therapeutic targets for the condition via the use of local network theoretic algorithms.

Our investigation was two-fold, we first considered the transcriptional landscape induced by high *DUX4* expression. We revealed that *DUX4* modified the expression of genes involved in apoptosis, oxidative stress, proliferation, uritogenital development and vasculature development. We also derived a *DUX4* gene expression signature, which validated in multiple independent data sets. We further demonstrated that *DUX4* induces a stem cell like phenotype characterised by an elevated signalling entropy. We next experimentally characterised then validated this inhibition of differentiation, in a well controlled cellular model of FSHD. We revealed by high-throughput time course image analysis that alignment and fusion of *DUX4* expressing, FSHD myoblasts is slower and less complete than their healthy counterparts.

Though there is a clear relationship between the *DUX4* over-expression and FSHD phenotypes, the low expression of *DUX4* in FSHD primary myoblasts led us to question whether there may be more useful targets for mitigating the pathology. To investigate this we applied the local analogue of signalling entropy, InSpiRe, introduced in detail in Chapter 2.

By applying InSpiRe to a meta-analysis of gene expression data, we constructed the FSHD network: the first unified, unbiased map of network rewiring underlying FSHD. Moreover, we revealed that the expression of genes in our FSHD network was significantly altered by *DUX4*, consistent with the growing consensus that FSHD is caused by aberrant *DUX4* expression [225, 388]. Differential gene expression studies have previously implicated disjointed collections of genes (*e.g.* myogenesis [213] and vasculature growth factors [187]), and hypotheses have been proposed on FSHD pathogenesis [186, 177]. However, there has been no unbiased, data driven insight into how these pathways relate, let alone actually co-ordinate. Our results emphasise the importance of development and application of network theoretic tools, such as InSpiRe, to identify pathomechanisms and therapeutics in complex diseases.

CTNNB1 has the highest betweenness centrality in our FSHD network, identifying β-catenin as a critical bottleneck of FSHD signalling. Gene expression correlation between β-catenin and its interaction partners is also significantly increased in FSHD, implying increased activation. Moreover, the TCF family of transcription factors are present in our network, as well as several upstream components of the canonical Wnt signalling pathway. Involvement of Wnt signalling in tissue specific myogenesis and in retinal angiogenesis led to a recent hypothesis that Wnt/β-catenin signalling was important for FSHD pathome-

chanisms [186]. A role of Wnt/β-catenin signalling in controlling *DUX4* expression and FSHD muscle cell apoptosis was also recently examined [389]. Moreover, enhancing the expression of β-catenin, via administration of $LiCl_2$ has recently been shown to reduce apoptosis in a model of occulopharyngeal muscular dystrophy [390]. Our FSHD network, compiled algorithmically from multiple independent data-sets, provides the detailed interactions of dysregulated β-catenin signalling. We also validated via qPCR that *DUX4* modifies the expression of critical downstream targets of β-catenin signalling.

The interaction partners of β-catenin in our FSHD network reveal mechanisms by which Wnt/β-catenin signalling is involved in certain hallmark phenotypes of FSHD through pathway cross-talk. Among these hallmarks is the characteristic oxidative stress sensitivity of FSHD muscle/myoblasts [235, 180]. Our results implicate HIF1-α signalling as critically perturbed in FSHD and show that *HIF1A* displays strong positive correlation with β-catenin in FSHD samples, providing mechanism for previous observations of the involvement of downstream components of the HIF1-α pathway in FSHD [52]. HIF1-α binds β-catenin during hypoxia competitively with TCF-4 [391], which may both inhibit cell proliferation activated by the TCF-4/β-catenin complex, as well as increase transcription of hypoxic response genes, including *VEGF*. We found that in FSHD, correlation in gene expression between β-catenin, and both *HIF1A* and TCF-4, significantly increases, resulting in elevated angiogenic genes such as *VEGF*, providing an explanation for the hallmark oxidative stress sensitivity in FSHD, as well as retinal vasculature abnormalities [186, 187].

Actin cytoskeleletal signalling has been implicated in FSHD pathology [377]. Our work shows perturbed cross-talk between such signalling and HIF1-α may contribute to FSHD. *MAP2* is associated with microtubule stability and rewired in FSHD. In hypoxic cardiomyocytes, *MAP2* is required for stabilisation of the microtubule network, leading to suppression of pVHL and increased HIF1-α [392]. Interestingly via this mechanism HIF1-α was up-regulated at the protein level but not the mRNA level (in early stages of hypoxia) [392], emphasising the power of our method for detecting events invisible to differential expression analysis of microarrays.

The role of ROS in FSHD is well reported, and our FSHD network contains many genes in the ROS-mediated pathway, including TNF-α. *DUX4* represses genes of the glutathione redox pathway, likely causing ROS accumulation in FSHD muscle [175], which may lead to increased TNF-α as part of a pro-inflammatory response [393]. Additionally, levels of TNF-α are negatively correlated with muscle endurance [180]. FSHD myoblasts undergo cell death in response to non-pathological levels of hydrogen peroxide [235] and other oxidative stress inducing factors; while antioxidants inhibit *DUX4*-induced toxicity in

FSHD myoblasts [175]. Thus over-activation of ROS-mediated TNF-α induced cell death pathways are potentially important pathomechanism in FSHD. TNF-α also stimulates ROS production via interaction with NADPH-oxidase, in a positive feedback loop [387]. Our results support this occurring in FSHD as *RAC1* is a well-connected member of our FSHD network and a critical component of the NADPH complex, also capable of activating JNK cell death signalling. *RAC1* is also activated by non-canonical Wnt signalling in a manner dependant on *MAP4K5* [394]. Our network unifies TNF-α, Wnt and JNK signalling in FSHD.

JNK signalling also plays an important role in oxidative stress-induced cell death, and we found that JNK signalling is more active in FSHD. JNK signalling can be activated in many ways, including via TNF-α signalling. In this scenario the AP-1 transcription factor genes *JUN*, *JUNB* and *FOS* are specifically up-regulated downstream of JNK. All these genes are present in our FSHD network, moreover, all display a significant increase in positive gene expression correlation with one another in FSHD samples, indicating their co-regulation. This result is evidence of TNF-α mediated JNK signalling in FSHD.

In the absence of NF-κB activity, prolonged JNK activation by TNF-α leads to apoptosis [395]. Due to the increased apoptosis of FSHD muscle cells, we postulate that NF-κB may be less active. In addition to activating JNK cell death pathways in response to ROS, TNF-α also activates NF-κB survival signalling causing the production of the anti-oxidant MnSOD to suppress ROS and minimise JNK cell death signalling [387]. MnSOD is the only anti-oxidant down regulated in FSHD [180], suggesting that NF-κB may be less active in FSHD muscle. Our results provide evidence for this theory: *NFKB1* encodes the DNA binding subunit of the NF-κB transcription factor, which though present in our FSHD network, is relatively uncorrelated with its interaction partners in FSHD samples. This implies *NFKB1* is inactive in FSHD muscle, and thus unable to repress the increased cell death via over-active JNK. Finally, we find crosstalk between JNK signalling and Wnt, in that all members of the PAR-1 gene family are in our FSHD network. This family was identified as dishevelled kinases, capable of simultaneously regulating Wnt activation of β-catenin signalling and JNK signalling [396].

Overactive JNK signalling has also been implicated in sensorineural hearing loss, an FSHD clinical phenotype [397]. Inhibitors of JNK signalling could partially mitigate the oxidative stress sensitivity of FSHD muscle cells, and D-JNKI-1 is currently in clinical trials for treatment of strokes [398].

In conclusion, we have demonstrated via signalling entropy and microscopy that *DUX4* expression leads to inhibited myogenesis in FSHD. We have then employed InSpiRe, to detect network rewiring in FSHD, performing the first meta-analysis of the FSHD tran-

scriptome and creating an integrated FSHD network. We also analysed a microarray of *DUX4* driven transcriptional changes and demonstrated that expression of genes in our network is significantly altered by *DUX4*. Our FSHD network provides an unbiased, unifying molecular map of FSHD signalling, elucidating perturbed genes and pathways critical to pathomechanisms. Importantly, we identify and validate β-catenin as central to FSHD pathology. Our network provides insight into the crucial steps of dysregulated signalling in FSHD and so will inform design of well-targeted therapeutics: currently lacking for FSHD.

5.4 Materials and Methods, Chapter 5

5.4.1 Cell Culture

During proliferation, the immortalised human myoblast cell lines (54-6 and 54-12) were grown in Skeletal Muscle Cell Growth Medium (Promocel) supplemented with 20% Foetal Bovine Serum, 50μg/ml Fetuin (bovine), 10ng/ml Epidermal Growth Factor (recombinant human), 1ng/ml Basic Fibroblast Growth Factor (recombinant human), 10μg/ml Insulin (recombinant human), 0.4μg/ml Dexamethasone and 50μg/ml Gentamycin, at 37^{o}C under 5% CO_2.

5.4.2 Immunocytochemistry

Plated myoblast cultures were fixed in 4% paraformaldehyde, permeabilised in 0.5% triton-X100 (Sigma Aldrich) for 10 minutes, washed thrice with PBS then blocked for a further 30 minutes at room temperature in 10% swine serum (DAKO) and 10% goat serum (DAKO) diluted in PBS. The plated cultures were subsequently incubated on a rocker in primary antibodies diluted in PBS supplemented with 1% goat serum (DAKO) overnight at 4^{o}C. Samples were then incubated at room temperature with AlexaFluor conjugated secondary antibodies (Life technologies) diluted in PBS supplemented with 1% goat serum, for 30 minutes, to detect primary antibody binding, before incubation with 1:1000 DAPI diluted in PBS for 10 minutes at room temperature. Primary antibodies used in this study include: mouse anti-Tubulin (1:1000 dilution) and mouse anti-*MyHC* (MF20) (1:400 dilution).
Images were acquired on a Zeiss Axiovert 200 M microscope using a Zeiss AxioCam HRm and AxioVision software version 4.4 (Zeiss). At least 5 fields were taken at 10$\times$ magnification for each cell culture well, resulting in quantification of over 500 cells per well.

5.4.3 5-Ethynyl-2'-deoxyuridine (EdU) incorporation

Proliferating immortalised myoblast cell lines were plated 5×10^3 in 8-well chamber slides (nunc). Cells were maintained in proliferation media (see above) overnight and were then incubated in EdU (Life technologies) as per the manufacturer's instructions, for two hours duration, before fixation in 4% paraformaldehyde. Fixed myoblast cultures were immunostained for tubulin (see above) prior to EdU detection with AlexaFluor azide 594 (Life technologies) (as per the manufacturers protocol).

5.4.4 Oxidative Stress Sensitivity

Proliferating immortalised myoblast cell lines were plated 5×10^3 in 8-well chamber slides (nunc). Cells were maintained in proliferation media (see above) overnight and were then switched to either fresh proliferation media or proliferation media containing 400μM H_2O_2. Cells were cultured for 24 hours before fixation in 4% paraformaldehyde. Fixed myoblast cultures were immunostained for Tubulin (described above).

5.4.5 Cell Differentiation

For differentiation the lines were plated at confluency (to avoid fusion rates driven by cell density differences acquired by different proliferation rates between the lines) in serum free DMEM Glutamax, supplemented with 50μg/ml Gentamycin, 10μg/ml Insulin (bovine) and 100μg/ml Apotransferin (human) and cultured at 37^oC under 5% CO_2. For immunocytochemical analysis of fusion cells were plated at 2.5×10^4 per well in a 96 well plate and cultured in differentiation medium for 2, 3 or 5 days fixation in 4% paraformaldehyde. Fixed myoblast cultures were immunostained for *MyHC* (described above).

5.4.6 Time course imaging

Immortalised myoblast cell lines 54-6 and 54-12 were plated in triplicate at confluency at the centre of a 96 well plate (25×10^4 cells per well) and induced to differentiate with a high volume (350 μl) of differentiation media (to prevent media evaporation during the 5 day imaging from altering the differentiation process). All remaining empty wells in the 96 plate were filled with DMEM Glutamax to provide humidity to the culture chamber in which they will be incubated.

Immediately after the addition of differentiation media cells were placed into a Solent Scientific chamber at 37^oC and 5% CO_2. Cells were imaged using an Eclipse Ti-E Live Cell Imaging System by taking a $10\times$ magnification, phase contrast image every 5 minutes from each well over a total of 5 days; this generated 1440 images per well per cell line,

resulting in a total of 8640 images.

5.4.7 Image analysis

Given the high quantity of image based readouts from immunocytochemical assays and the time consuming nature of manual image analysis, it was important to obtain a high throughput, unbiased analytic method. We thus wrote a high throughput image analysis software in the statistical programming language R, using the image processing package EBImage [382]. The software can autonomously process hundreds of high quality, large images in the order of minutes, providing readings of cell counts in 3 colour channels as well as computing the fusion index for differentiated cell cultures. Figures derived from the software were confirmed representative of manually derived values. The EBImage package has extensive options for automated cell phenotyping and thus characterisation of cell morphology is also possible in a high throughput manner.

Briefly, each image is first split into 3 channels, for nuclear counting a low pass filter is used the remove noise and binarise the image, the binarised image is then filtered to remove objects which are not sufficient large and circular enough to be consider cell nuclei, finally any holes in the resulting image are filled. Particles in the binarised image are then quantified and output to a csv file.

For determination of fusion index, a low pass filter is used on the channel displaying *MyHC* to quantify the area positive for *MyHC* and binarise the image into *MyHC* positive and *MyHC* negative subsets, subsequently the DAPI channel is processed as above for cell counting and evaluated for total DAPI positive area. Regions in the processed DAPI image which are classified are *MyHC* negative are then set to zero, the DAPI positive *MyHC* positive area is then computed and divided by the total DAPI positive area to give the fusion index.

To perform cell phenotyping on proliferating cells stained with tubulin, a low pass filter is used on the tubulin channel, and the DAPI channel is processed for cell counting as described above. To avoid overlapping cell cytoplasm from skewing the automated cell phenotyping, cells are segmented using a Voroni based method, where DAPI defined nuclei are used as seeds [382]. Fig 5.20 presents a flow chart of the image analysis process.

Image analysis software to interrogate the myogenesis time-course imaging data simply adapted the immunocytochemistry software. Each image taken in the time course was passed through a low pass filter to reduce noise and binarise the image, high intensity regions were filtered on size and morphology to remove regions not considered likely to correspond to cells and finally holes were filled in. The regions of high intensity follow-

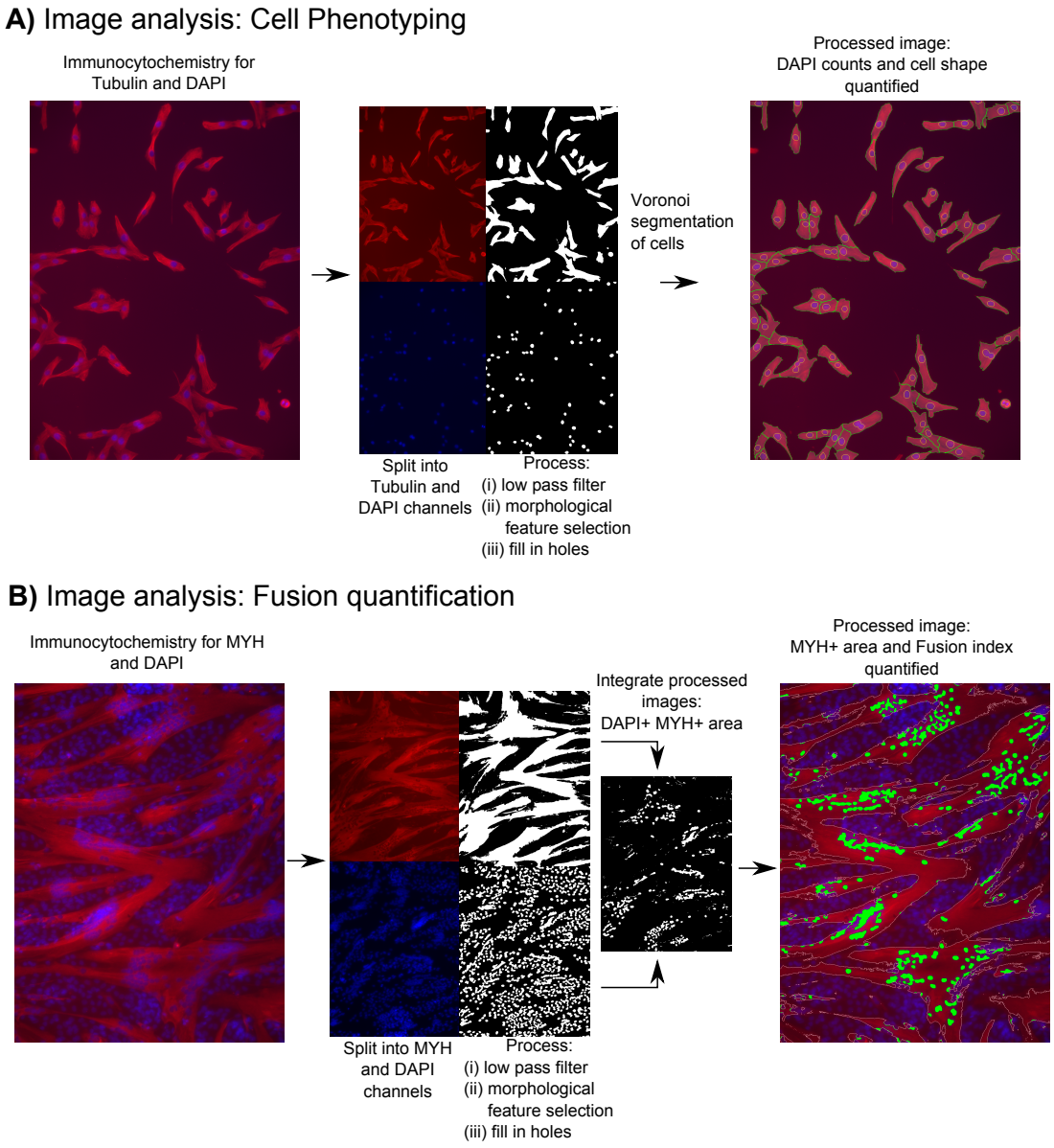

Figure 5.20: **Automated image analysis of immunocytochemistry.** (A) For cell counting and morphology analysis images were separated into colour channels and binarised via a low pass filter. Morphological filters were employed to remove background and a Voronoi segmentation anchored on DAPI positive nuclei was used to segment contacting cell cytoplasm, for morphological analysis of single cells. (B) To assess myotube positive area and fusion index, images were split into colour channels and binarised via a low pass filter. Areas of *MyHC* positive and DAPI positive area were computed and DAPI positive *MyHC* positive area was also computed and normalised by total DAPI positive area to give the fusion index.

ing this thresholding typically corresponded to single cells. The eccentricity of each cell identified in each image by this thresholding was measured and the average eccentricity obtained for each 5 minute time point. We thus obtained a time course of eccentricities in triplicate for each cell line.
Differential eccentricities between cell lines were assessed by an empirical Bayes approach and p-value histograms confirmed that differential eccentricities were detectable. Non-linear regression was utilised to fit polynomial curves to the mean (over the triplicates) eccentricity time course, for each cell line, the turning points of the curves were determined to assess the most extremely eccentric and non-eccentric time points for each cell line separately.

5.4.8 Harvesting cells for RNA

Immortalised myoblast cell lines 54-6 and 54-12 were plated in triplicate at confluency in 12 well plates (3.12×10^5 cells per well) and induced to differentiate with 1ml of differentiation media. At each selected time point for transcriptomic analysis cells were imaged on a light microscope, before being washed with 500μl PBS and incubated with 150μl trypsin at 37^oC for 5 minutes. Trypsin was inactivated and cells were suspended in 1ml warm proliferation media, before cells were pelleted by centrifugation at 4,500 rpm for 5 minutes. The supernatant was carefully removed by pipetting and cells were stored at -80^oC, until all samples were collected and RNA extraction could commence.

5.4.9 Signalling Entropy

Signalling entropy was computed as described in Chapter 1.

5.4.10 Expression Data sets

From the GEO Database [270], two studies were sufficiently large for reliable inference, GSE3307 [399] (14 FSHD samples, 14 controls) and GSE10760 [187] (19 FSHD samples 30 controls), both profiled on the Affymetrix Human Genome U133A Array platform. An additional recent large study available at GEO accession number GSE36398, was disregarded due to substantial discordance with other studies in terms of signal intensity [377]. Exon array studies of FSHD and control myoblasts and myotubes, GSE26145 [52] and GSE26061 [233] were also considered. Both studies employed similar experimental conditions, each containing 3 samples of each cell type and were profiled on the Affymetrix Human Exon 1.0 ST Array. Importantly, isolated cell types were the source of mRNA so

noise from non-muscle specific gene expression is removed.
We also obtained from the GEO database human skeletal muscle studies into gene expression in muscle diseases other than FSHD (GSE3307 [399]), age (GSE5086 [400] and GSE9676 [401]) and atrophy (GSE5110 [402] and GSE8872 [403]).
All expression data sets considered underwent a quality control step using the ArrayQualityMetrics package in R [404]. This step included analysis of RNA degradation and elimination of samples displaying clear signal saturation effects [405]. Data sets were log normalised using RMA and PCA was performed to determine if the dominant principal component correlated with condition being compared (*e.g.* disease status). PCA performed on the two larger FSHD studies (GSE3307 and GSE10760) revealed that the dominant principal component correlated with both disease status and age of patients sampled. For this reason we additionally analysed data sets associated with age dependant gene expression.
PCA on the age dependant study, GSE5086 revealed that the dominant principal component correlated with age and seperated the samples into 2 groups corresponding to young and older samples. We used this clustering to classify the samples in this data set as either younger or older patients. Non-FSHD muscle diseases analysed were selected from GSE3307 due to the large number of high quality samples available.
GSE2614 and GSE26061 were suitable for integration by the ComBat function [378] in R which employs an empirical Bayes approach to eliminate batch effects, and was recently demonstrated as superior to a wide range of other microarray data integration methods [406]. The integrated data set contained 6 FSHD myoblasts, 6 FSHD myotubes, 6 control myobalsts and 6 control myotubes. PCA was performed on the integrated data set and the dominant principal component correlated with cell type and disease status but not with batch, implying a successful integration.
We extracted from data sets probes mapping to genes corresponding to proteins in our interaction network. For probes mapping to the same gene we computed the average expression across these probes for each sample, assigning the resulting value to the gene. For each data set, proteins in the interaction network with no corresponding probe in the microarrays were deleted from the network, proteins with a degree of zero following this deletion were also removed from the network. This resulted in a reduced PIN for each data set.

5.4.11 Construction of the PIN

The human PIN utilised in this section was constructed as described in Chapter 2. The mouse PIN was constructed from the maximal connected component of our human PIN,

after vertices were mapped from human to mouse by orthololgy relations.

5.4.12 The InSpiRe Algorithm

The InSpiRe algorithm was implemented as described in detail in Chapter 2.

5.4.13 Comparing Possible Transformed Pearson Correlations in Step 1 of InSpiRe

The choice of the transformation of the Pearson correlations to construct the stochastic matrix required for the InSpiRe algorithm described in Chapter 2 must ensure non-negativity, and that the interpretation of a high edge weight indicative of an increased likelihood of interaction between connected proteins is valid. As discussed we considered two possible transformations:

$$w_{ij} = \frac{1}{2}(1 + C_{ij}) \tag{85}$$

and

$$w_{ij} = |C_{ij}| \tag{86}$$

where C_{ij} denotes the Pearson correlation in the gene expression profiles of protein i and protein j across samples corresponding to the phenotype considered. Equation (85) assumes that signal transduction flows preferably along paths of proteins with positively correlated gene expression profiles and has been employed previously [23, 17]. It was noted, however, that this was only an approximation [23], and there is mounting evidence that negative correlations play an important role in signal transduction [250], consequentially the transformation described by equation (86), which assigns both strong positive and strong negative correlations a high weight may be considered more realistic for this interpretation.

We utilised both transformations independently and upon performing step 2 of InSpiRe to identify proteins significantly rewiring between FSHD and control skeletal muscle, we found that the transformation (86) provided a substantially better discriminator than (85) as judged by p-value histograms [407]. In fact, all p-value histograms utilising the transformation described in (85) were flat, implying that in utilising this transformation one is unable to reliably determine differences between the phenotypes. Consequentially all results described in this chapter were obtained utilising the more realistic transformation (86).

5.4.14 Comparing methodologies

To evaluate the performance of InSpiRe relative to other methodologies, we applied InSpiRe, NetWalk [16] and GSEA [368] on differentially expressed genes to each FSHD data set independently, and evaluated the enrichments of identified genes.

5.4.14.1 Differential expression analysis Differential expression analysis was performed on normalised data sets matched to the protein interaction network using the limma package in R [269]. GSEA [368] was then performed against the gene sets of the Molecular Signatures Database [369], using the t-scores output by the limma analysis to rank the genes. Gene sets identified by GSEA as displaying $p < 0.05$ and $FDR < 0.25$ were considered significantly enriched.

5.4.14.2 NetWalk analysis NetWalk is a network based algorithm described in Chapter 1, which considers the stationary distribution of a weighted random walk on a network of compiled interactions. Weights on network vertices are data derived and bias walker visitation in a biologically relevant manner. We implemented NetWalk on normalised data sets using the NetWalker software [408], and employing the compiled Knowledgebase provided as the underlying network for implementation; functional annotation of identified edges was performed using the FunWalk option. The i^{th} element of the weight flux vector, $(\mathbf{w})_{i=1}^{N}$, where N is the number of genes in the network, for a given data set, was defined as the ratio of the mean expression the i^{th} probe across samples corresponding to the phenotype examined (disease, aged, atrophic) to the mean expression across control samples. This selection is a recommended option [408]. In order to ensure the findings of NetWalk were statistically robust, we utilised the jackknife re-sampling procedure to assess significance of edge visitation ratios, and functional annotation ratios. As with InSpiRe, NetWalk was run independently on each data set to produce a list of significant functional terms, and each FSHD data set identified around 3000 significant functional terms. The intersection of the significant functional terms in the FSHD data set consisted of 266 terms, and contained several terms associated to oxidative stress, apoptosis and mitochondrial dysfunction. When terms also associated with age, atrophy and other diseases were removed from this intersection however, only 19 terms remained, none of which had a strong justification to association with FSHD in the literature.

5.4.14.3 Functional annotation for InSpiRe implicated genes Functional annotation of InSpiRe implicated genes was performed on the significant genes ($p < 0.05$)

implicated by local flux entropy or local symmetrised KL divergence, using the DAVID Bioinformatics Resources 6.7 [308]. The p-value cut off for the Fisher's exact test (EASE score) employed by the DAVID software for implicating enriched pathways was 0.05. The background gene set utilised consisted of all the genes in the PIN.

5.4.14.4 Comparison To compare the relevance of the results output by the various methodologies, we considered 14 FSHD associated pathways: Wnt signalling, *TNF* or *MAPK* related signalling, vasculature development, calcium signalling, oxidative stress response, cell cycle, apoptosis, mitochondrial dysfunction, asymmetrical development, muscle structure, nuclear envelope, muscle differentiation, histone modification and actin cytoskeletal signalling. For each pathway we gave each method a score from 0 to 4 corresponding to the number of FSHD data sets it was capable of detecting the pathway in. Of the three methodologies considered InSpiRe was the most successful, achieving an average score of 3.5, NetWalk achieved 3.29 and GSEA on differentially expressed genes achieved 1.86.

5.4.15 Re-sampling procedure to assess concordance between microarray and the the FSHD network

To determine whether expression of genes in our FSHD network was significantly altered by *DUX4*, we first identified a probeset of mouse orthologs to the genes in the FSHD network consisting of 1866 genes, to permit comparison with our murine satellite cell microarray of *DUX4* construct expression. Our objective was to assess whether the expression of the genes in the network probeset was able to distinguish between the 5 *DUX4* retroviral constructs better than would be expected by chance. If this is the case, then the clustering of the *DUX4* constructs based on the expression of the network probeset should be significantly better than that based on a random probeset of equivalent size.
We therefore performed a re-sampling procedure, evaluating 10000 random probesets of 1866 genes from our microarray. For each random probeset we performed a hierarchical clustering and enforced a 6 cluster solution, which we then compared to the optimal 6 cluster solution corresponding to the perfect separation of the 6 retroviral constructs. To compare cluster distributions we used the Rand index, a function which assesses classification similarity on a unit scale. In this manner we obtained a null distribution of Rand indices describing how well a random probeset may be expected to cluster the 6 *DUX4* constructs. This null distribution then allowed us to calculate a p-value evaluating the hypothesis that the expression of the network probeset clustered the *DUX4* constructs

better than could be expected by chance.

5.4.16 qPCR

For quantification of *DUX4* over expressing murine satellite cell-derived myoblasts, cells were infected with *DUX4* and control retroviral constructs and RNA extracted at 24 hours and 48 hours after infection using Qiagen RNeasy Kit and quantification on a Nanodrop ND-1000 spectrophotometer (Labtech). For quantification of low levels of *DUX4* in immortalised human myoblasts, RNA integrity was also quantified on an Aligent Bioanalyzer.
RNA was then reverse-transcribed using the Reverse Transcription Kit with genomic DNA wipeout (Qiagen) and qPCR was performed on an Mx3005P qPCR system (Stratagene) with MESA Blue qPCR MasterMix Plus and ROX reference dye (Eurogentec). Primers used were as follows:
Lef1; F- TCATCACCTACAGCGACGAG, R-TGATGGGAAAACCTGGACAT. *Lgr5*; F- CCGCCAGTCTCCTACATCGCC, R- GCATTGTCATCTAGCCACAGGTGCC. *Lgr6*; F- CACACATCCCGGGACAGGCAT, R- GGGAGGAGAGCCCCTCAAGC. *Tcf3*; F- TCTCAAGCCGGTTCCCACAC, R- TTTCCGGGCAAGCTCATAGTATTT. *Tcf4*; F- TGCCGACTACAACAGGGACT, R- TGCTGGACTGTGGGATATGA. *Myf5* F- TGAGGGAACAGGTGGAGAAC, R-AGCTGGACACGGAGCTTTTA; *DUX4* F- 5'-CCCAGGTACCAGCAGACC-3' R- 5'-TCCAGGAGATGTAACTCTAATCCA-3'.

6 Overview, Discussion and Philosophy

6.1 On our philosophy of approach

In this thesis we have considered the interplay between mathematics and biology, in an attempt to elucidate global principles and local mechanisms underlying certain biological phenomena and complex pathologies. The motivation underlying this work came from many sources, but fundamentally it derives from the sheer complexity of the biological system, which results in the generation of data sets whose analysis is inconceivable without the application of mathematical methodologies.
We began this work with an exploration of network theory in biology because it is clear that a graphical representation of intra-cellular interactions represents a information rich, minimal parameter description of the system. Given the nature of the broad but shallow data typically produced in the biological sciences such a representation seems well suited

to investigating what we can currently observe. Moreover, the description is flexible and can be readily adapted as the experimenters eyes adjust through new technology.

We saw how the early forays into network biology revealed novel insights into the nature of intra-cellular interactions. In particular we noted the discovery of network rewiring, in which the true biological response can only be found in differential interactions. In our expose on current network rewiring approaches, however, we presented a concern, which to my eyes applies to many aspects of high-throughput, data-driven mathematical biology. Much work in this field focuses on the encapsulation of pathways and genes which may be attributed a particular process or pathology. For example, some transcriptomic signatures associate with survival in breast cancer and some genes involved in myogenesis are differentially expressed in FSHD. These findings are important and they provide us with information about features we should target in these pathologies. What do we learn, however, about the nature of pathological development? What do we understand about the general dynamics of the biological system? To me developing sophisticated approaches for the listing of these differences, without interrogating the general principles behind why they may indeed be different, is a suboptimal strategy. It tailors our understanding of processes to specific data sets, given scenarios which may not scale across the population. Moreover, it gives us little insight into novel processes, which can only be understood laboriously, by repetition of the same data-driven approach.

To my eyes an important part of a mathematician's role in any applied science is the positing of global laws governing the behaviour of systems under study. If these laws validate in observation, they tell us something about the nature of the system, rather than the nature of the data. Our goal in this work was therefore to consider network based mathematical methodologies, which can form the basis of a hypothesis driven approach to biological data analysis. Such methodologies must yield a theoretical framework, permitting one to posit and examine certain global principles, as well as being applicable to current data, permitting validation. Moreover, the methods should be developed in a scalable manner, and hence, once validated, can generate the list of differential features characteristic of other methodologies. In this way we do not lose what can be gained from the differential approach, rather we build on it, obtaining an intuitive understanding of the system.

The positing of global laws underlying biological network rewiring is fundamentally a postulate of permissible trajectories in a high dimensional space. More precisely, the state of a biological network at any given moment, can be considered in terms of elements of one or more metric spaces. The rewiring of a this network from one state to another, can thus be described as a trajectory through this space. The nature of the network

and the process undertaken during state transition, necessarily imposes restrictions on the possible trajectories. It is the clarification of these restrictions that will elucidate our global laws of biological network rewiring. If a suitable framework can be constructed, it provides the basis to evaluate the merits of a given trajectory restriction hypothesis. If validated these hypotheses can be used to build up contours in our metric spaces, firming up the foundations of our theoretical framework.

Such hypotheses are best first generated by consideration of observable properties of biological systems, evolving during processes where network rewiring is known to take place. In the introduction we consider cellular differentiation as precisely such a process. This is a truly systems level process, evolving as a consequence of coordinated rewiring. We saw that as cells differentiate, it has been observed that they homogenise their expression profiles, though the individual gene expression patterns show great diversity across lineages. We hypothesised therefore, that a restriction on the network rewiring trajectory from an undifferentiated to a differentiated biological state, will be a reduction in the disorder of network interactions.

In truth, this hypothesis, its validation and its consequences in pathology has been at the heart of all the work undertaken in this thesis.

6.2 An overview of our work

We began our work, as seemed natural, with the consideration of metric spaces. Here we abstracted our notion of a network state to a combination of structure, defined by a weighted graph, and dynamics, as defined by probability distributions for vertex interactions. We saw how the two notions were coupled and that deformation of network structure influenced dynamics naturally within our framework across all possible state transition trajectories. Moreover, we saw how this framework could be utilised to consider information transfer within a given state via NTE, and hence to compare information transfer across network states. Thus our framework permits the phenotype comparisons typical of conventional biological data analyses.

We progressed to provide a theoretical basis for the evaluation of our hypothesis on permissible network rewiring trajectories during cellular differentiation. We considered the entropy rate of our network dynamics model, dubbed signalling entropy, and demonstrated that it was a population measure of heterogeneity across biological cells.

Following the generation of theoretical constructs, we considered large amounts of experimental data as a means to validate our hypothesis that signalling entropy decreased throughout cellular differentiation. We found that our hypothesis was indeed correct in over 1000 samples corresponding to cells at different stages of the differentiation hierarchy.

Moreover, we found that as differentiation progressed signalling entropy dropped systematically, even reversing when conditions were manipulated to allow de-differentiation.
We thus elucidated a restriction on biologically permissible network rewiring trajectories during cellular differentiation. As this process progresses signalling entropy decreases.
Given such a trajectory restriction in cellular differentiation, it is only natural to consider states where this process is faulty, namely pathologies of development. What happens to signalling entropy in such pathologies? How is it different to the healthy case? What can this tell us about the developmental defect? What are the genes and proteins that are driving this? These are all important questions which can elucidate therapeutic targets for disease and crucially describe global properties underlying them.
We therefore considered our measure in two antithetical diseases of development: breast cancer and FSHD.
For cancer we saw that signalling entropy is elevated in CSCs and is powerfully prognostic regardless of subtype and across multiple malignancies. This is a result worthy of contemplation. There currently exist both histological and transcriptomic prognostic indicators for a range of malignancies, moreover, there exist many biomarkers for CSCs arising in different tissues. The features which define these indicators and biomarkers are, however, specific to a given tissue, or even to a single cancer subtype; they do not show significant overlap; they are derived from data sets not global principles. Is it surprising that their discriminatory power is limited to the setting from which they were derived, whilst that of signalling entropy is not? To me this is a clear validation of the importance of global principles in biology. Why waste time developing many tissue specific cancer prognostic indicators, when from global considerations rooted in theory, we can develop a single indicator which is as good as, or better than those derived from data?
For FSHD, we saw that expression of the primary candidate gene *DUX4* induces an elevated signalling entropy in muscle cell precursors. We showed by detailed characterisation of muscle differentiation that FSHD myoblasts undergo a slower, inhibited muscle differentiation in line with this result. However, the network theoretic methods so readily applied in cancer are lacking in FSHD, with little known about druggable targets and pathways. We thus considered a local analogue of signalling entropy, the InSpiRe algorithm. Performing a meta-analysis of FSHD muscle biopsy, gene expression data sets, we uncovered a network describing FSHD pathomechanisms and detailed a multitude of potential targets for investigation. We found that β-catenin was central to FSHD and validated its perturbation in cellular models.
We thus, in this thesis, considered a hypothesis regarding a *simple* restriction on a permissible network rewiring trajectory during a *single* biological process. Yet we have seen

that the applications to health and pathology are diverse, and we regret that we have only scratched their surface in this work.

6.3 Future Directions

Much remains to be done. This is not to say that the work of this thesis is incomplete, rather it is impress a notion of the size of the field that can be approached whilst building on this work.

Firstly, we can continue to develop the notion of signalling entropy as a measure of cell potency, theoretically there are number of questions which currently spring to mind: What sort of gene expression regimes are required to increase or decrease signalling entropy? How much do these depend on the starting state of the network? What is the minimal number of genes we must modify to change signalling entropy in a predictable way? These questions can all be addressed in the framework of control theory, and can of course be coupled with experimental validations using set-ups like siRNA mediated gene knock-down and retroviral mediated gene over-expression. It is important to understand the controllability of signalling entropy given its relation to cancer prognostics and disease gene mediated phenotypes. Control theory derived strategies may form the basis of powerful therapeutic regimes, most likely in cancer.

From a biological point of view, much remains to be investigated regarding signalling entropy associations with pathological processes. There is a dearth of time course gene expression data describing pathological cell differentiation, but such data represents a powerful resource for investigating signalling entropy as a measure of cell potency in a pathological context. Though we have seen that signalling entropy correlates with tumour stemness and is elevated by *DUX4* expression, how this elevation comes about from the healthy state is unclear. Time course data of pathological differentiation and tumour progression, therefore may elucidate network rewiring trajectories which are followed during the descent into pathology, thus providing the means for their reversion.

The nature of signalling entropy as a measure of cell potency in a healthy context is currently attributed to a mixture of intra-cellular signalling promiscuity and inter-cellular heterogeneity. However, which of these factors is dominant is unclear. The generation of single cell, genome wide, gene expression data is thus essential for the investigation of this concept. Moreover, the application of signalling entropy to the guided differentiation of pluripotent stem cells into desired lineages for regenerative medicine and iPSC derived pathological models, merits a detailed investigation.

We only scratched the surface of signalling entropy's association with tissue specific cancer

mortality, and this is something that must be investigated in more detail. Limitations in our medical technologies, such as targeted therapies and early detection thorough symptom presentation and imaging, are thought to underlie much of the discrepancies in tissue specific cancer mortality. The fact that signalling entropy associates with this variable, however, suggests that there is something deeper at play and if we can address it we may be able to normalise survival across cancer types.

In FSHD, we uncovered a myriad of pathogenic mechanisms which merit investigation. Arguably the most promising are β-catenin signalling and HIF1-α mediated oxidative stress sensitivity. The targeting of these pathways may yield the first viable therapeutics for this highly prevalent pathology.

It is of course clear that in addition to FSHD and cancer, there are a seemingly endless stream of development related pathologies, many of which may benefit from the application of signalling entropy to an understanding of their pathogenesis. Hence the application of the techniques developed in this thesis to other diseases can be considered an important topic of future work.

It must further be emphasised that the investigations of signalling entropy in this thesis, represent the exploration of a single hypothesis on the restriction of a network rewiring trajectory during a *single* biological process. There are a huge number of other biological processes, each with their own associated systems level properties, validation data sets and related pathologies *e.g.,* proliferation, apoptosis, necrosis, activation (say of T-cells), migration (chemotaxis), quiescence, phagocytosis, *etc.*. For each of these processes another thesis can be written, positing and validating a restriction on a network rewiring trajectory and investigating the impact of that restriction on associated pathology. Such restrictions can be combined with signalling entropy to further elucidate the landscape of permissive biological network rewiring.

Finally our metric space framework, which can be considered analogous to network rewiring state space requires more theoretical development. We outlined in Chapter 1 a few of the possible avenues this can take, such as the investigation of network evolution by dynamic programming and the investigation of persistent network states by vertex cycle decompositions. Arguably many more investigations are possible and must be considered an important topic of future work.

7 Abbreviations

Shorthand	Longhand
ARACNe	Algorithm for the Reconstruction of Accurate Cellular Networks
ATRA	all-trans retinoic acid
BANJO	Bayesian Network Inference with Java Objects
ChIP	chromatin immunoprecipitation
CNV	copy number variation
Co-IP	co-immunoprecipitation
COX	cytochrome c oxidase
CSC	cancer stem cell
CTC	circulating tumour cell
DAPI	4',6-diamidino-2-phenylindole
DMSO	dimethylsulphoxide
DPI	Data Processing Inequality
EdU	5-Ethynyl-2'-deoxyuridine
EGF	epidermal growth factor
EMT	epithelial to mesenchymal transition
eQED	Expression Quantitative trait loci Electrical Diagrams
ER	oestrogen receptor
ES	embryonic stem
FACS	fluorescence-activated cell sorting
FSHD	facioscapulohumeral muscular dystrophy
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
H&E	haematoxylin and eosin
HPV	human papilloma virus
HR	Hormone receptor
HSC	hematopoietic stem cell
IDC-NST	invasive ductal carcinomas of no special type
InSpiRe	Interactome Sparsification and Rewiring
iPSC	induced pluripotent stem cell
ISD	initial signal distribution
JNK	c-Jun N-terminal kinase
JSD	Jensen-Shannon divergence
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium

NIR	Network Inference by Reverse-engineering
NSC	neural stem cell
NSCLC	non-small cell lung cancer
NTE	Network Transfer Entropy
ODE	ordinary differential equation
PCA	principal component analysis
PIN	protein interaction network
PR	progesterone receptor
qPCR	quantitative reverse transcription polymerase chain reaction
RNA-seq	RNA-sequencing
ROS	reactive oxygen species
RPE	retinal pigment epithelium
SCC	squamous cell carcinoma
SCM	Stem Cell Matrix
SE Score	signalling entropy prognostic score
SNP	single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
TPE	telomere position effect
WHO	World Health Organisation
Y2H	yeast 2 hybrid

8 References

[1] Le Fanu, J. *The rise and fall of modern medicine* (Abacus, London, 2011), fully rev. and updated [ed.]. edn.

[2] Wang, E. *Cancer Systems Biology.* Chapman & Hall/CRC Mathematical & Computational Biology (CRC Press, 2010).

[3] Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–13 (2004).

[4] Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & di Bernardo, D. How to infer gene networks from expression profiles. *Mol Syst Biol* **3**, 78 (2007).

[5] Margolin, A. A. *et al.* Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7 (2006).

[6] Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J. & Jarvis, E. D. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **20**, 3594–603 (2004).

[7] Gregoretti, F., Belcastro, V., di Bernardo, D. & Oliva, G. A parallel implementation of the network identification by multiple regression (nir) algorithm to reverse-engineer regulatory gene networks. *PLoS One* **5**, e10179 (2010).

[8] Cerami, E. G. *et al.* Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, D685–90 (2011).

[9] Chatr-Aryamontri, A. *et al.* The biogrid interaction database: 2013 update. *Nucleic Acids Res* **41**, D816–23 (2013).

[10] Goel, R., Harsha, H. C., Pandey, A. & Prasad, T. S. Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. *Mol Biosyst* **8**, 453–63 (2012).

[11] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**, D109–14 (2012).

[12] Kerrien, S. *et al.* The intact molecular interaction database in 2012. *Nucleic Acids Res* **40**, D841–6 (2012).

[13] Bollobas, B. *Modern graph theory*, vol. 184 (Springer, 1998).

[14] Serrano, M. A., Boguna, M. & Sagues, F. Uncovering the hidden geometry behind metabolic networks. *Mol Biosyst* **8**, 843–50 (2012).

[15] Kim, Y. A., Wuchty, S. & Przytycka, T. M. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol* **7**, e1001095 (2011).

[16] Komurov, K., White, M. A. & Ram, P. T. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput Biol* **6** (2010).

[17] Teschendorff, A. E. & Severini, S. Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst Biol* **4**, 104 (2010).

[18] Roy, J., Winter, C., Isik, Z. & Schroeder, M. Network information improves cancer outcome prediction. *Brief Bioinform* (2012).

[19] Winter, C. *et al.* Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* **8**, e1002511 (2012).

[20] Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to dna damage. *Science* **330**, 1385–9 (2010).

[21] Macarthur, B. D., Ma'ayan, A. & Lemischka, I. R. Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol* **10**, 672–81 (2009).

[22] Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol Syst Biol* **5**, 321 (2009).

[23] West, J., Bianconi, G., Severini, S. & Teschendorff, A. E. Differential network entropy reveals cancer system hallmarks. *Sci Rep* **2**, 802 (2012).

[24] Sambrook, J. & Russell, D. W. Identification of associated proteins by coimmunoprecipitation. *CSH Protoc* **2006** (2006).

[25] Bruckner, A., Polge, C., Lentze, N., Auerbach, D. & Schlattner, U. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci* **10**, 2763–88 (2009).

[26] Komurov, K. & Ram, P. T. Patterns of human gene expression variance show strong associations with signaling network hierarchy. *BMC Syst Biol* **4**, 154 (2010).

[27] de la Fuente, A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet* **26**, 326–33 (2010).

[28] Lee, W. P. & Tzou, W. S. Computational methods for discovering gene networks from expression data. *Brief Bioinform* **10**, 408–23 (2009).

[29] Hartemink, A. J. Reverse engineering gene regulatory networks. *Nat Biotechnol* **23**, 554–5 (2005).

[30] Beirlant, J., Dudewicz, E. J., Gyorfi, L. & Van der Meulen, E. C. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences* **6**, 17–40 (1997).

[31] Basso, K. *et al.* Reverse engineering of regulatory networks in human b cells. *Nat Genet* **37**, 382–90 (2005).

[32] Hecker, M., Lambeck, S., Toepfer, S., van Someren, E. & Guthke, R. Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems* **96**, 86–103 (2009).

[33] Jensen, S. T., Chen, G. & Stoeckert Jr, C. J. Bayesian variable selection and data integration for biological regulatory networks. *The Annals of Applied Statistics* 612–633 (2007).

[34] Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–12 (1999).

[35] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–4 (2000).

[36] Albert, R., Jeong, H. & Barabasi, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378–82 (2000).

[37] Han, J. D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).

[38] Stumpf, M. P., Wiuf, C. & May, R. M. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci U S A* **102**, 4221–4 (2005).

[39] Khanin, R. & Wit, E. How scale-free are biological networks. *J Comput Biol* **13**, 810–8 (2006).

[40] Winterbach, W., Van Mieghem, P., Reinders, M., Wang, H. & de Ridder, D. Topology of molecular interaction networks. *BMC Syst Biol* **7**, 90 (2013).

[41] Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nat Biotechnol* **18**, 1257–61 (2000).

[42] Ma, H. W., Zhao, X. M., Yuan, Y. J. & Zeng, A. P. Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* **20**, 1870–6 (2004).

[43] Newman, M. E. J. A measure of betweenness centrality based on random walks. *Social Networks* **27**, 39–54 (2005).

[44] Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* **27**, 199–204 (2009).

[45] Serrano, M. A., Krioukov, D. & Boguna, M. Self-similarity of complex networks and hidden metric spaces. *Phys Rev Lett* **100**, 078701 (2008).

[46] Califano, A. Rewiring makes the difference. *Mol Syst Biol* **7**, 463 (2011).

[47] Ideker, T. & Krogan, N. J. Differential network biology. *Mol Syst Biol* **8**, 565 (2012).

[48] Mentzen, W. I., Floris, M. & de la Fuente, A. Dissecting the dynamics of dysregulation of cellular processes in mouse mammary gland tumor. *BMC Genomics* **10**, 601 (2009).

[49] Lai, Y., Wu, B., Chen, L. & Zhao, H. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* **20**, 3146–55 (2004).

[50] Suthram, S., Beyer, A., Karp, R. M., Eldar, Y. & Ideker, T. eqed: an efficient method for interpreting eqtl associations using protein networks. *Mol Syst Biol* **4**, 162 (2008).

[51] Ben-David, U. & Benvenisty, N. The tumorigenicity of human embryonic and induced pluripotent stem cells. *Nat Rev Cancer* **11**, 268–77 (2011).

[52] Tsumagari, K. *et al.* Gene expression during normal and FSHD myogenesis. *BMC Med Genomics* **4**, 67 (2011).

[53] Waddington, C. H. *The strategy of the genes : a discussion of some aspects of theoretical biology* (Allen & Unwin, London, 1957).

[54] Macarthur, B. D. & Lemischka, I. R. Statistical mechanics of pluripotency. *Cell* **154**, 484–489 (2013).

[55] Merkle, F. T. & Eggan, K. Modeling human disease with pluripotent stem cells: from genome association to function. *Cell Stem Cell* **12**, 656–68 (2013).

[56] Cohen, D. E. & Melton, D. Turning straw into gold: directing cell fate for regenerative medicine. *Nat Rev Genet* **12**, 243–52 (2011).

[57] Pardal, R., Clarke, M. F. & Morrison, S. J. Applying the principles of stem-cell biology to cancer. *Nat Rev Cancer* **3**, 895–902 (2003).

[58] Rosner, M. & Hengstschlager, M. Intercellular protein expression variability as a feature of stem cell pluripotency. *Amino Acids* (2013).

[59] Muller, F. J. *et al.* A bioinformatic assay for pluripotency in human cells. *Nat Methods* **8**, 315–7 (2011).

[60] Nadig, R. R. Stem cell therapy - hype or hope? a review. *J Conserv Dent* **12**, 131–8 (2009).

[61] Segers, V. F. & Lee, R. T. Stem-cell therapy for cardiac disease. *Nature* **451**, 937–42 (2008).

[62] Knoepfler, P. S. Deconstructing stem cell tumorigenicity: a roadmap to safe regenerative medicine. *Stem Cells* **27**, 1050–6 (2009).

[63] Marjanovic, N. D., Weinberg, R. A. & Chaffer, C. L. Cell plasticity and heterogeneity in cancer. *Clin Chem* **59**, 168–79 (2013).

[64] Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).

[65] Robertson, N. J. *et al.* Embryonic stem cell-derived tissues are immunogenic but their inherent immune privilege promotes the induction of tolerance. *Proc Natl Acad Sci U S A* **104**, 20920–5 (2007).

[66] Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–76 (2006).

[67] Guha, P., Morgan, J. W., Mostoslavsky, G., Rodrigues, N. P. & Boyd, A. S. Lack of immune response to differentiated cells derived from syngeneic induced pluripotent stem cells. *Cell Stem Cell* **12**, 407–12 (2013).

[68] Boyd, A. S., Rodrigues, N. P., Lui, K. O., Fu, X. & Xu, Y. Concise review: Immune recognition of induced pluripotent stem cells. *Stem Cells* **30**, 797–803 (2012).

[69] Cahan, P. & Daley, G. Q. Origins and implications of pluripotent stem cell variability and heterogeneity. *Nat Rev Mol Cell Biol* **14**, 357–68 (2013).

[70] Gutteridge, A. *et al.* Novel pancreatic endocrine maturation pathways identified by genomic profiling and causal reasoning. *PLoS One* **8**, e56024 (2013).

[71] Park, I. H. *et al.* Disease-specific induced pluripotent stem cells. *Cell* **134**, 877–86 (2008).

[72] Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R. C. & Melton, D. A. "stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298**, 597–600 (2002).

[73] Sato, N. *et al.* Molecular signature of human embryonic stem cells and its comparison with the mouse. *Dev Biol* **260**, 404–13 (2003).

[74] Evsikov, A. V. & Solter, D. Comment on " 'stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* **302**, 393; author reply 393 (2003).

[75] Fortunel, N. O. *et al.* Comment on " 'stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* **302**, 393; author reply 393 (2003).

[76] Chen, L. & Daley, G. Q. Molecular basis of pluripotency. *Hum Mol Genet* **17**, R23–7 (2008).

[77] Chambers, I. *et al.* Functional expression cloning of nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113**, 643–55 (2003).

[78] Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364–8 (2006).

[79] F. J. Mueller, e. Regulatory networks define phenotypic classes of human stem cell lines. *Nature* **455**, 401–405 (2008).

[80] Furusawa, C. & Kaneko, K. A dynamical-systems view of stem cell biology. *Science* **338**, 215–7 (2012).

[81] Wang, J., Xu, L., Wang, E. & Huang, S. The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophys J* **99**, 29–39 (2010).

[82] Enver, T., Pera, M., Peterson, C. & Andrews, P. W. Stem cell states, fates, and the rules of attraction. *Cell Stem Cell* **4**, 387–97 (2009).

[83] Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D. E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys Rev Lett* **94**, 128701 (2005).

[84] Mar, J. C. & Quackenbush, J. Decomposition of gene expression state space trajectories. *PLoS Comput Biol* **5**, e1000626 (2009).

[85] Ladewig, J., Koch, P. & Brustle, O. Leveling waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nat Rev Mol Cell Biol* **14**, 225–36 (2013).

[86] Zhou, J. X., Brusch, L. & Huang, S. Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model. *PLoS One* **6**, e14752 (2011).

[87] Zhou, J. X. & Huang, S. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends Genet* **27**, 55–62 (2011).

[88] Huang, S., Guo, Y. P., May, G. & Enver, T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol* **305**, 695–713 (2007).

[89] Heinaniemi, M. *et al.* Gene-pair expression signatures reveal lineage control. *Nat Methods* **10**, 577–83 (2013).

[90] Wang, J., Zhang, K., Xu, L. & Wang, E. Quantifying the waddington landscape and biological paths for development and differentiation. *Proc Natl Acad Sci U S A* **108**, 8257–62 (2011).

[91] Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–6 (2002).

[92] Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415–8 (2004).

[93] Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A* **98**, 8614–9 (2001).

[94] Till, J. E., McCulloch, E. A. & Siminovitch, L. A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proc Natl Acad Sci U S A* **51**, 29–36 (1964).

[95] Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* **6**, 451–64 (2005).

[96] Simpson, P. Notch signalling in development: on equivalence groups and asymmetric developmental potential. *Curr Opin Genet Dev* **7**, 537–42 (1997).

[97] Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453**, 544–7 (2008).

[98] Furusawa, C. & Kaneko, K. Chaotic expression dynamics implies pluripotency: when theory and experiment meet. *Biol Direct* **4**, 17 (2009).

[99] ONS. Cancer statistics registrations. *England (Series MB1)* **No. 42** (2011).

[100] Lyratzopoulos, G. & Abel, G. Earlier diagnosis of breast cancer: focusing on symptomatic women. *Nat Rev Clin Oncol* **10**, 544 (2013).

[101] Marmot, M. G. *et al.* The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* **108**, 2205–40 (2013).

[102] Autier, P., Boniol, M. & Boyle, P. The benefits and harms of breast cancer screening. *Lancet* **381**, 800 (2013).

[103] De Abreu, F. B., Wells, W. A. & Tsongalis, G. J. The emerging role of the molecular diagnostics laboratory in breast cancer personalized medicine. *Am J Pathol* (2013).

[104] Berrington de Gonzalez, A. & Reeves, G. Mammographic screening before age 50 years in the uk: comparison of the radiation risks with the mortality benefits. *Br J Cancer* **93**, 590–6 (2005).

[105] Jalalian, A. *et al.* Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review. *Clin Imaging* **37**, 420–6 (2013).

[106] Boyd, N. F. Mammographic density and risk of breast cancer. *Am Soc Clin Oncol Educ Book* **2013**, 57–62 (2013).

[107] Elston, C. W. & Ellis, I. O. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* **19**, 403–10 (1991).

[108] Weigelt, B. & Reis-Filho, J. S. Histological and molecular types of breast cancer: is there a unifying taxonomy? *Nat Rev Clin Oncol* **6**, 718–30 (2009).

[109] Galea, M. H., Blamey, R. W., Elston, C. E. & Ellis, I. O. The nottingham prognostic index in primary breast cancer. *Breast Cancer Res Treat* **22**, 207–19 (1992).

[110] Weinberg, R. A. *The biology of Cancer*, vol. 255 (Garland Science, New York, 2007), 1 edn.

[111] Lakhani, S., Ellis, I., Schnitt, S., Tan, P. & van de Vijver, M. *WHO Classification of Tumours of the Breast* (International Agency for Research on Cancer, 2012), 4 edn.

[112] Ellis, I. O. *et al.* Pathological prognostic factors in breast cancer. ii. histological type. relationship with survival in a large study with long-term follow-up. *Histopathology* **20**, 479–89 (1992).

[113] Rakha, E. A. *et al.* Morphological and immunophenotypic analysis of breast carcinomas with basal and myoepithelial differentiation. *J Pathol* **208**, 495–506 (2006).

[114] Mathieu, M. C. *et al.* The poor responsiveness of infiltrating lobular breast carcinomas to neoadjuvant chemotherapy can be explained by their biological profile. *Eur J Cancer* **40**, 342–51 (2004).

[115] Beatson, G. On the treatment of inoperable cases of carcinoma of the mamma: Suggestions for a new method of treatment, with illustrative cases. *Lancet* **148**, 162–165 (1896).

[116] McGuire, W. L. Estrogen receptors in human breast cancer. *J Clin Invest* **52**, 73–7 (1973).

[117] Wolff, A. C. & Dowsett, M. Estrogen receptor: a never ending story? *J Clin Oncol* **29**, 2955–8 (2011).

[118] Ignatiadis, M. & Sotiriou, C. Luminal breast cancer: from biology to treatment. *Nat Rev Clin Oncol* **10**, 494–506 (2013).

[119] Lim, E., Metzger-Filho, O. & Winer, E. P. The natural history of hormone receptor-positive breast cancer. *Oncology (Williston Park)* **26**, 688–94, 696 (2012).

[120] Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–93 (2012).

[121] Jordan, V. C. Tamoxifen: a most unlikely pioneering medicine. *Nat Rev Drug Discov* **2**, 205–13 (2003).

[122] Sree, S. V., Ng, E. Y., Acharya, R. U. & Faust, O. Breast imaging: A survey. *World J Clin Oncol* **2**, 171–8 (2011).

[123] Musgrove, E. A. & Sutherland, R. L. Biological determinants of endocrine resistance in breast cancer. *Nat Rev Cancer* **9**, 631–43 (2009).

[124] Brisken, C. Progesterone signalling in breast cancer: a neglected hormone coming into the limelight. *Nat Rev Cancer* **13**, 385–96 (2013).

[125] Acharya, U. R., Ng, E. Y., Tan, J. H. & Sree, S. V. Thermography based breast cancer detection using texture features and support vector machine. *J Med Syst* **36**, 1503–10 (2012).

[126] Skaane, P. Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review. *Acta Radiol* **50**, 3–14 (2009).

[127] Burstein, H. J. The distinctive nature of her2-positive breast cancers. *N Engl J Med* **353**, 1652–4 (2005).

[128] Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290**, 261–4 (1981).

[129] Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *Science* **235**, 177–82 (1987).

[130] Arteaga, C. L. *et al.* Treatment of her2-positive breast cancer: current status and future perspectives. *Nat Rev Clin Oncol* **9**, 16–32 (2012).

[131] Valabrega, G., Montemurro, F. & Aglietta, M. Trastuzumab: mechanism of action, resistance and future perspectives in her2-overexpressing breast cancer. *Ann Oncol* **18**, 977–84 (2007).

[132] Spector, N. L. *et al.* Study of the biologic effects of lapatinib, a reversible inhibitor of erbb1 and erbb2 tyrosine kinases, on tumor growth and survival pathways in patients with advanced malignancies. *J Clin Oncol* **23**, 2502–12 (2005).

[133] Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* **121**, 2750–67 (2011).

[134] Engebraaten, O., Vollan, H. K. & Borresen-Dale, A. L. Triple-negative breast cancer and the need for new therapeutic targets. *Am J Pathol* (2013).

[135] Martins, W. K. A useful procedure to isolate simultaneously dna and rna from a single tumor sample (2009).

[136] Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–52 (2000).

[137] Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7**, 96 (2006).

[138] Eroles, P., Bosch, A., Perez-Fidalgo, J. A. & Lluch, A. Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. *Cancer Treat Rev* **38**, 698–707 (2012).

[139] Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol* **8**, R76 (2007).

[140] Peto, R. *et al.* Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet* **379**, 432–44 (2012).

[141] Paik, S. *et al.* Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J Clin Oncol* **24**, 3726–34 (2006).

[142] Small, G. W., Shi, Y. Y., Higgins, L. S. & Orlowski, R. Z. Mitogen-activated protein kinase phosphatase-1 is a mediator of breast cancer chemoresistance. *Cancer Res* **67**, 4459–66 (2007).

[143] Ellis, M. J. *et al.* Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* **486**, 353–60 (2012).

[144] Ithimakin, S. *et al.* Her2 drives luminal breast cancer stem cells in the absence of her2 amplification: implications for efficacy of adjuvant trastuzumab. *Cancer Res* **73**, 1635–46 (2013).

[145] Staaf, J. *et al.* Identification of subtypes in human epidermal growth factor receptor 2–positive breast cancer reveals a gene signature prognostic of outcome. *J Clin Oncol* **28**, 1813–20 (2010).

[146] Xia, W. *et al.* A model of acquired autoresistance to a potent erbb2 tyrosine kinase inhibitor and a therapeutic strategy to prevent its onset in breast cancer. *Proc Natl Acad Sci U S A* **103**, 7795–800 (2006).

[147] Moll, R., Franke, W. W., Schiller, D. L., Geiger, B. & Krepler, R. The catalog of human cytokeratins: patterns of expression in normal epithelia, tumors and cultured cells. *Cell* **31**, 11–24 (1982).

[148] Gusterson, B. Do 'basal-like' breast cancers really exist? *Nat Rev Cancer* **9**, 128–34 (2009).

[149] Turner, N. C. & Reis-Filho, J. S. Basal-like breast cancer and the brca1 phenotype. *Oncogene* **25**, 5846–53 (2006).

[150] Subhawong, A. P. *et al.* Most basal-like breast carcinomas demonstrate the same rb-/p16+ immunophenotype as the hpv-related poorly differentiated squamous cell

carcinomas which they resemble morphologically. *Am J Surg Pathol* **33**, 163–75 (2009).

[151] Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418–23 (2003).

[152] Lee, E. *et al.* Characteristics of triple-negative breast cancer in patients with a brca1 mutation: results from a population-based study of young women. *J Clin Oncol* **29**, 4373–80 (2011).

[153] Badve, S. *et al.* Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Mod Pathol* **24**, 157–67 (2011).

[154] Helleday, T. The underlying mechanism for the parp and brca synthetic lethality: clearing up the misunderstandings. *Mol Oncol* **5**, 387–93 (2011).

[155] Creighton, C. J. *et al.* Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proc Natl Acad Sci U S A* **106**, 13820–5 (2009).

[156] Perou, C. M. Molecular stratification of triple-negative breast cancers. *Oncologist* **16 Suppl 1**, 61–70 (2011).

[157] Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–52 (2012).

[158] van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999–2009 (2002).

[159] Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817–26 (2004).

[160] Nguyen, L. V., Vanner, R., Dirks, P. & Eaves, C. J. Cancer stem cells: an evolving concept. *Nat Rev Cancer* **12**, 133–43 (2012).

[161] Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–11 (2001).

[162] Stingl, J. & Caldas, C. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nat Rev Cancer* **7**, 791–9 (2007).

[163] Booth, B. W. & Smith, G. H. Estrogen receptor-alpha and progesterone receptor are expressed in label-retaining mammary epithelial cells that divide asymmetrically and retain their template dna strands. *Breast Cancer Res* **8**, R49 (2006).

[164] Shackleton, M., Quintana, E., Fearon, E. R. & Morrison, S. J. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell* **138**, 822–9 (2009).

[165] Pinto, C. A., Widodo, E., Waltham, M. & Thompson, E. W. Breast cancer stem cells and epithelial mesenchymal plasticity - implications for chemoresistance. *Cancer Lett* **341**, 56–62 (2013).

[166] McDermott, S. P. & Wicha, M. S. Targeting breast cancer stem cells. *Mol Oncol* **4**, 404–19 (2010).

[167] Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci U S A* **100**, 3983–8 (2003).

[168] Owens, T. W. & Naylor, M. J. Breast cancer stem cells. *Front Physiol* **4**, 225 (2013).

[169] Quintana, E. *et al.* Efficient tumour formation by single human melanoma cells. *Nature* **456**, 593–8 (2008).

[170] Nakshatri, H., Srour, E. F. & Badve, S. Breast cancer stem cells and intrinsic subtypes: controversies rage on. *Curr Stem Cell Res Ther* **4**, 50–60 (2009).

[171] Chaffer, C. L. *et al.* Poised chromatin at the zeb1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. *Cell* **154**, 61–74 (2013).

[172] Emery, A. E. The muscular dystrophies. *Lancet* **359**, 687–95 (2002).

[173] Scott, W., Stevens, J. & Binder-Macleod, S. A. Human skeletal muscle fiber type classifications. *Phys Ther* **81**, 1810–6 (2001).

[174] Relaix, F. & Zammit, P. S. Satellite cells are essential for skeletal muscle regeneration: the cell on the edge returns centre stage. *Development* **139**, 2845–56 (2012).

[175] Bosnakovski, D. *et al.* An isogenetic myoblast expression screen identifies dux4-mediated fshd-associated molecular pathologies. *EMBO J* **27**, 2766–79 (2008).

[176] Orrell, R. W. Facioscapulohumeral dystrophy and scapuloperoneal syndromes. *Handb Clin Neurol* **101**, 167–80 (2011).

[177] Lefkowitz, D. L. & Lefkowitz, S. S. Fascioscapulohumeral muscular dystrophy: a progressive degenerative disease that responds to diltiazem. *Med Hypotheses* **65**, 716–21 (2005).

[178] Orrell, R. W., Copeland, S. & Rose, M. R. Scapular fixation in muscular dystrophy. *Cochrane Database Syst Rev* CD003278 (2010).

[179] Rose, M. R. & Tawil, R. Drug treatment for facioscapulohumeral muscular dystrophy. *Cochrane Database Syst Rev* CD002276 (2004).

[180] Turki, A. *et al.* Functional muscle impairment in facioscapulohumeral muscular dystrophy is correlated with oxidative stress and mitochondrial dysfunction. *Free Radic Biol Med* **53**, 1068–79 (2012).

[181] Wallace, L. M. *et al.* Dux4, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy in vivo. *Ann Neurol* **69**, 540–52 (2011).

[182] Pandya, S., King, W. M. & Tawil, R. Facioscapulohumeral dystrophy. *Phys Ther* **88**, 105–13 (2008).

[183] Padberg, G. W. *Facioscapulohumeral Disease*. Ph.D. thesis (1982).

[184] Cooper, D. & Upadhhyaya, M. *Facioscapulohumeral Muscular Dystrophy (FSHD): Clinical Medicine and Molecular Cell Biology* (Taylor & Francis, 2004).

[185] Desai, U. R. & Sabates, F. N. Long-term follow-up of facioscapulohumeral muscular dystrophy and coats' disease. *Am J Ophthalmol* **110**, 568–9 (1990).

[186] Fitzsimons, R. B. Retinal vascular disease and the pathogenesis of facioscapulohumeral muscular dystrophy. A signalling message from Wnt? *Neuromuscul Disord* **21**, 263–71 (2011).

[187] Osborne, R. J., Welle, S., Venance, S. L., Thornton, C. A. & Tawil, R. Expression profile of fshd supports a link between retinal vasculopathy and muscular dystrophy. *Neurology* **68**, 569–77 (2007).

[188] Brouwer, O. F. *et al.* Hearing loss in facioscapulohumeral muscular dystrophy. *Neurology* **41**, 1878–81 (1991).

[189] Lutz, K. L., Holte, L., Kliethermes, S. A., Stephan, C. & Mathews, K. D. Clinical and genetic features of hearing loss in facioscapulohumeral muscular dystrophy. *Neurology* **81**, 1374–7 (2013).

[190] Trevisan, C. P. *et al.* Facioscapulohumeral muscular dystrophy: a multicenter study on hearing function. *Audiol Neurootol* **13**, 1–6 (2008).

[191] Trevisan, C. P. *et al.* Facioscapulohumeral muscular dystrophy and occurrence of heart arrhythmia. *Eur Neurol* **56**, 1–5 (2006).

[192] Landouzy, J., L.; Dejerine. *De la myopathie atrophique progressive (myopathie hereditaire debutant, dans l'enfance, par la face, sans alteration du systeme nerveux* (Gauthier-Villars, Paris, 1884).

[193] Tyler, F. H. & Stephens, F. E. Studies in disorders of muscle. ii clinical manifestations and inheritance of facioscapulohumeral dystrophy in a large family. *Ann Intern Med* **32**, 640–60 (1950).

[194] Kazakov, V., Rudenko, D., Skorometz, A. & Kolynin, V. Scapuloperoneal muscular dystrophy is an independent variant of fshd? *Acta Myol* **28**, 103 (2009).

[195] Padberg, G. W. *et al.* On the significance of retinal vascular disease and hearing loss in facioscapulohumeral muscular dystrophy. *Muscle Nerve Suppl* S73–80 (1995).

[196] Statland, J. M. *et al.* Coats syndrome in facioscapulohumeral dystrophy type 1: frequency and d4z4 contraction size. *Neurology* **80**, 1247–50 (2013).

[197] Fitzsimons, R. B., Gurwin, E. B. & Bird, A. C. Retinal vascular abnormalities in facioscapulohumeral muscular dystrophy. a general association with genetic and therapeutic implications. *Brain* **110 (Pt 3)**, 631–48 (1987).

[198] Halpin, C., Owen, G., Gutierrez-Espeleta, G. A., Sims, K. & Rehm, H. L. Audiologic features of norrie disease. *Ann Otol Rhinol Laryngol* **114**, 533–8 (2005).

[199] Bass, S. J., Sherman, J. & Giovinazzo, V. Bilateral coats' response in a female patient leads to diagnosis of facioscapulohumeral muscular dystrophy. *Optometry* **82**, 72–6 (2011).

[200] Tsuji, M., Kinoshita, M., Imai, Y., Kawamoto, M. & Kohara, N. Facioscapulohumeral muscular dystrophy presenting with hypertrophic cardiomyopathy: a case study. *Neuromuscul Disord* **19**, 140–2 (2009).

[201] Della Marca, G. *et al.* Heart rate variability in facioscapulohumeral muscular dystrophy. *Funct Neurol* **25**, 211–6 (2010).

[202] Galetta, F. *et al.* Subclinical cardiac involvement in patients with facioscapulohumeral muscular dystrophy. *Neuromuscul Disord* **15**, 403–8 (2005).

[203] Wijmenga, C. *et al.* Mapping of facioscapulohumeral muscular dystrophy gene to chromosome 4q35-qter by multipoint linkage analysis and in situ hybridization. *Genomics* **9**, 570–5 (1991).

[204] van Deutekom, J. C. *et al.* Fshd associated dna rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum Mol Genet* **2**, 2037–42 (1993).

[205] Tawil, R. *et al.* Evidence for anticipation and association of deletion size with severity in facioscapulohumeral muscular dystrophy. the fsh-dy group. *Ann Neurol* **39**, 744–8 (1996).

[206] Tupler, R. *et al.* Monosomy of distal 4q does not cause facioscapulohumeral muscular dystrophy. *J Med Genet* **33**, 366–70 (1996).

[207] de Greef, J. C. *et al.* Clinical features of facioscapulohumeral muscular dystrophy 2. *Neurology* **75**, 1548–54 (2010).

[208] de Greef, J. C. *et al.* Common epigenetic changes of d4z4 in contraction-dependent and contraction-independent fshd. *Hum Mutat* **30**, 1449–59 (2009).

[209] Hartweck, L. M. *et al.* A focal domain of extreme demethylation within d4z4 in fshd2. *Neurology* **80**, 392–9 (2013).

[210] Zeng, W. *et al.* Specific loss of histone h3 lysine 9 trimethylation and hp1gamma/cohesin binding at d4z4 repeats is associated with facioscapulohumeral dystrophy (fshd). *PLoS Genet* **5**, e1000559 (2009).

[211] Gabellini, D. *et al.* Facioscapulohumeral muscular dystrophy in mice overexpressing frg1. *Nature* **439**, 973–7 (2006).

[212] Pistoni, M. *et al.* Rbfox1 downregulation and altered calpain 3 splicing by frg1 in a mouse model of facioscapulohumeral muscular dystrophy (fshd). *PLoS Genet* **9**, e1003186 (2013).

[213] Winokur, S. T. *et al.* Expression profiling of fshd muscle supports a defect in specific stages of myogenic differentiation. *Hum Mol Genet* **12**, 2895–907 (2003).

[214] Gabriels, J. *et al.* Nucleotide sequence of the partially deleted d4z4 locus in a patient with fshd identifies a putative gene within each 3.3 kb element. *Gene* **236**, 25–32 (1999).

[215] Kawamura-Saito, M. *et al.* Fusion between cic and dux4 up-regulates pea3 family genes in ewing-like sarcomas with t(4;19)(q35;q13) translocation. *Hum Mol Genet* **15**, 2125–37 (2006).

[216] Snider, L. *et al.* Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *PLoS Genet* **6**, e1001181 (2010).

[217] Ansseau, E. *et al.* Dux4c is up-regulated in fshd. it induces the myf5 protein and human myoblast proliferation. *PLoS One* **4**, e7482 (2009).

[218] Bosnakovski, D. *et al.* Dux4c, an fshd candidate gene, interferes with myogenic regulators and abolishes myoblast differentiation. *Exp Neurol* **214**, 87–96 (2008).

[219] Leidenroth, A. & Hewitt, J. E. A family history of dux4: phylogenetic analysis of duxa, b, c and duxbl reveals the ancestral dux gene. *BMC Evol Biol* **10**, 364 (2010).

[220] Bosnakovski, D., Daughters, R. S., Xu, Z., Slack, J. M. & Kyba, M. Biphasic myopathic phenotype of mouse dux, an orf within conserved fshd-related repeats. *PLoS One* **4**, e7003 (2009).

[221] Clapp, J. *et al.* Evolutionary conservation of a coding function for D4Z4, the tandem DNA repeat mutated in facioscapulohumeral muscular dystrophy. *Am J Hum Genet* **81**, 264–79 (2007).

[222] Dixit, M. *et al.* Dux4, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of pitx1. *Proc Natl Acad Sci U S A* **104**, 18157–62 (2007).

[223] Pandey, S. N. *et al.* Conditional over-expression of pitx1 causes skeletal muscle dystrophy in mice. *Biol Open* **1**, 629–639 (2012).

[224] Lemmers, R. J. *et al.* A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science* **329**, 1650–3 (2010).

[225] Lemmers, R. J. *et al.* Digenic inheritance of an smchd1 mutation and an fshd-permissive d4z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet* **44**, 1370–4 (2012).

[226] Dandapat, A., Hartweck, L. M., Bosnakovski, D. & Kyba, M. Expression of the human fshd-linked dux4 gene induces neurogenesis during differentiation of murine embryonic stem cells. *Stem Cells Dev* **22**, 2440–8 (2013).

[227] Tassin, A. *et al.* Dux4 expression in fshd muscle cells: how could such a rare protein cause a myopathy? *J Cell Mol Med* **17**, 76–89 (2013).

[228] Ehrlich, M. & Lacey, M. Deciphering transcription dysregulation in fsh muscular dystrophy. *J Hum Genet* **57**, 477–84 (2012).

[229] Scionti, I. *et al.* Large-scale population analysis challenges the current criteria for the molecular diagnosis of fascioscapulohumeral muscular dystrophy. *Am J Hum Genet* **90**, 628–35 (2012).

[230] Jones, T. I. *et al.* Facioscapulohumeral muscular dystrophy family studies of dux4 expression: evidence for disease modifiers and a quantitative model of pathogenesis. *Hum Mol Genet* **21**, 4419–30 (2012).

[231] Stadler, G. *et al.* Telomere position effect regulates dux4 in human facioscapulohumeral muscular dystrophy. *Nat Struct Mol Biol* **20**, 671–8 (2013).

[232] Baur, J. A., Zou, Y., Shay, J. W. & Wright, W. E. Telomere position effect in human cells. *Science* **292**, 2075–7 (2001).

[233] Cheli, S. *et al.* Expression profiling of FSHD-1 and FSHD-2 cells during myogenic differentiation evidences common and distinctive gene dysregulation patterns. *PLoS One* **6**, e20966 (2011).

[234] Geng, L. N. *et al.* Dux4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev Cell* **22**, 38–51 (2012).

[235] Barro, M. *et al.* Myoblasts from affected and non-affected fshd muscles exhibit morphological differentiation defects. *J Cell Mol Med* **14**, 275–89 (2010).

[236] Tassin, A. *et al.* Fshd myotubes with different phenotypes exhibit distinct proteomes. *PLoS One* **7**, e51865 (2012).

[237] Macaione, V. *et al.* RAGE-NF-kappaB pathway activation in response to oxidative stress in facioscapulohumeral muscular dystrophy. *Acta Neurol Scand* **115**, 115–21 (2007).

[238] Krom, Y. D. *et al.* Intrinsic epigenetic regulation of the d4z4 macrosatellite repeat in a transgenic mouse model for fshd. *PLoS Genet* **9**, e1003415 (2013).

[239] Copeland, S. A., Levy, O., Warner, G. C. & Dodenhoff, R. M. The shoulder in patients with muscular dystrophy. *Clin Orthop Relat Res* 80–91 (1999).

[240] Diab, M., Darras, B. T. & Shapiro, F. Scapulothoracic fusion for facioscapulohumeral muscular dystrophy. *J Bone Joint Surg Am* **87**, 2267–75 (2005).

[241] Krom, Y. D. *et al.* Generation of isogenic d4z4 contracted and noncontracted immortal muscle cell clones from a mosaic patient: a cellular model for fshd. *Am J Pathol* **181**, 1387–401 (2012).

[242] Ahuja, R. K., Magnanti, T. L. & Orlin, J. B. Network flows: theory, algorithms, and applications (1993).

[243] Baptiste, P., Le Pape, C. & Nuijten, W. *Constraint-based scheduling: applying constraint programming to scheduling problems*, vol. 39 (Springer, 2001).

[244] Stehlé, J., Barrat, A. & Bianconi, G. Dynamical and bursty interactions in social networks. *Physical review E* **81**, 035101 (2010).

[245] Paulsson, J. Models of stochastic gene expression. *Physics of life reviews* **2**, 157–175 (2005).

[246] Chen, H. & Mandelbaum, A. Discrete flow networks: bottleneck analysis and fluid approximations. *Mathematics of Operations Research* **16**, 408–446 (1991).

[247] Lin, Y.-K. A simple algorithm for reliability evaluation of a stochastic-flow network with node failure. *Computers & Operations Research* **28**, 1277–1285 (2001).

[248] Serfozo, R. *Introduction to stochastic networks*, vol. 44 (Springer, 1999).

[249] Gomez-Gardenes, J. & Latora, V. Entropy rate of diffusion processes on complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **78**, 065102 (2008).

[250] Zeng, T. & Li, J. Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucleic Acids Res* **38**, e1 (2010).

[251] Schreiber, T. Measuring information transfer. *Phys Rev Lett* **85**, 461–4 (2000).

[252] Endres, D. M. & Schindelin, J. E. A new metric for probability distributions. *Information Theory, IEEE Transactions on* **49**, 1858–1860 (2003).

[253] Csermely, P., Korcsmáros, T., Kiss, H. J., London, G. & Nussinov, R. Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. *Pharmacology & therapeutics* **138**, 333–408 (2013).

[254] Van Kampen, N. G. *Stochastic processes in physics and chemistry*, vol. 1 (Access Online via Elsevier, 1992).

[255] Vicente, R., Wibral, M., Lindner, M. & Pipa, G. Transfer entropy–a model-free measure of effective connectivity for the neurosciences. *J Comput Neurosci* **30**, 45–67 (2011).

[256] Kwon, O. & Yang, J.-S. Information flow between stock indices. *EPL (Europhysics Letters)* **82**, 68003 (2008).

[257] Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

[258] Wootters, W. K. Statistical distance and hilbert space. *Physical Review D* **23**, 357 (1981).

[259] Casas, M., Lamberti, P., Plastino, A. & Plastino, A. Jensen-shannon divergence, fisher information, and wootters' hypothesis. *arXiv preprint quant-ph/0407147* (2004).

[260] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–9 (2005).

[261] Bondy, J. A. & Murty, U. Graph theory, volume 244 of. *Graduate texts in mathematics* (2008).

[262] West, J. *Network Physics in Cancer and Aging.* Ph.D. thesis (2013).

[263] Uhlen, M. *et al.* Proteomics. tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

[264] Lovász, L. Random walks on graphs: A survey. *Combinatorics* **2**, 1–46 (1993).

[265] Ge, X. *et al.* Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* **86**, 127–41 (2005).

[266] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–8 (1976).

[267] van Wieringen, W. N. & van der Vaart, A. W. Statistical analysis of the cancer cell's molecular entropy using high-throughput data. *Bioinformatics* **27**, 556–63 (2011).

[268] Banerji, C. R., Severini, S. & Teschendorff, A. E. Network transfer entropy and metric space for causality inference. *Phys Rev E Stat Nonlin Soft Matter Phys* **87**, 052814 (2013).

[269] Smyth, G. K. *et al.* Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, 3 (2004).

[270] Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* **41**, D991–5 (2013).

[271] Prasad, T. S., Kandasamy, K. & Pandey, A. Human protein reference database and human proteinpedia as discovery tools for systems biology. *Methods Mol Biol* **577**, 67–79 (2009).

[272] Kandasamy, K. *et al.* Netpath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).

[273] Grimmett, G. & Stirzaker, D. *Probability and random processes* (Oxford University Press, 1992).

[274] Demetrius, L. & Manke, T. Robustness and network evolution-an entropic principle. *Physica A: Statistical Mechanics and its Applications* **346**, 682–696 (2005).

[275] T. S. Mikkelsen, e. Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49–55 (2008).

[276] K. L. Nazor, e. Recurrent variations in dna methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell* **10**, 620–634 (2012).

[277] Ben-Porath, I. *et al.* An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* **40**, 499–507 (2008).

[278] N. A. Watkins, e. A haematlas: characterizing gene expression in differentiated human blood cells. *Blood* **113**, e1–e9 (2009).

[279] Goldfarb, A. N., Wong, D. & Racke, F. K. Induction of megakaryocytic differentiation in primary human erythroblasts: a physiological basis for leukemic lineage plasticity. *Am J Pathol* **158**, 1191–1198 (2001).

[280] Miranda-Saavedra, D. & Gottgens, B. Transcriptional regulatory networks in haematopoiesis. *Curr Opin Genet Dev* **18**, 530–535 (2008).

[281] Yu, Y. H. *et al.* Network biology of tumor stem-like cells identified a regulatory role of CBX5 in lung cancer. *Sci Rep* **2**, 584 (2012).

[282] Noggle, S. A., Weiler, D. & Condie, B. G. Notch signaling is inactive but inducible in human embryonic stem cells. *Stem Cells* **24**, 1646–1653 (2006).

[283] K. Yu, e. A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet* **4**, e1000129 (2008).

[284] Meier-Stiegen, F. e. Activated notch1 target genes during embryonic cell differentiation depend on the cellular context and include lineage determinants and inhibitors. *PLoS One* **5**, e11481 (2010).

[285] Liu, J., Sato, C., Cerletti, M. & Wagers, A. Notch signaling in the regulation of stem cell self-renewal and differentiation. *Curr Top Dev Biol* **92**, 367–409 (2010).

[286] Bigas, A., D'Altri, T. & Espinosa, L. The notch pathway in hematopoietic stem cells. *Curr Top Microbiol Immunol* **360**, 1–18 (2012).

[287] Blank, U., Karlsson, G. & Karlsson, S. Signaling pathways governing stem-cell fate. *Blood* **111**, 492–503 (2008).

[288] M. Minami, e. Stat3 activation is a critical step in gp130-mediated terminal differentiation and growth arrest of a myeloid cell line. *Proc Natl Acad Sci U S A* **93**, 3963–3966 (1996).

[289] E. Caldenhoven, e. Differential activation of functionally distinct stat5 proteins by il-5 and gm-csf during eosinophil and neutrophil differentiation from human cd34+ hematopoietic stem cells. *Stem Cells* **16**, 397–403 (1998).

[290] Kanayasu-Toyoda, T., Yamaguchi, T., Uchida, E. & Hayakawa, T. Commitment of neutrophilic differentiation and proliferation of hl-60 cells coincides with expression of transferrin receptor. effect of granulocyte colony stimulating factor on differentiation and proliferation. *J Biol Chem* **274**, 25471–25480 (1999).

[291] Coffer, P. J., Koenderman, L. & de Groot, R. P. The role of stats in myeloid differentiation and leukemia. *Oncogene* **19**, 2511–2522 (2000).

[292] Palmer, N. P., Schmid, P. R., Berger, B. & Kohane, I. S. A gene expression profile of stem cell pluripotentiality and differentiation is conserved across diverse solid and hematopoietic cancers. *Genome Biol* **13**, R71 (2012).

[293] Banerji, C. R. S. *et al.* Cellular network entropy as the energy potential in waddington's differentiation landscape. *Sci Rep* **3**, 3039 (2013).

[294] Barberi, T., Willis, L. M., Socci, N. D. & Studer, L. Derivation of multipotent mesenchymal precursors from human embryonic stem cells. *PLoS Med* **2**, e161 (2005).

[295] Giraud-Triboult, K. *et al.* Combined mRNA and microRNA profiling reveals that miR-148a and miR-20b control human mesenchymal stem cell phenotype via EPAS1. *Physiol Genomics* **43**, 77–86 (2011).

[296] Tanabe, S. *et al.* Gene expression profiling of human mesenchymal stem cells for identification of novel markers in early- and late-stage cell culture. *J Biochem* **144**, 399–408 (2008).

[297] Wagner, W. *et al.* Replicative senescence of mesenchymal stem cells: a continuous and organized process. *PLoS One* **3**, e2213 (2008).

[298] Larson, B. L., Ylostalo, J. & Prockop, D. J. Human multipotent stromal cells undergo sharp transition from division to development in culture. *Stem Cells* **26**, 193–201 (2008).

[299] Avery, K., Avery, S., Shepherd, J., Heath, P. R. & Moore, H. Sphingosine-1-phosphate mediates transcriptional regulation of key targets associated with survival, proliferation, and pluripotency in human embryonic stem cells. *Stem Cells Dev* **17**, 1195–205 (2008).

[300] Ebert, A. D. *et al.* Induced pluripotent stem cells from a spinal muscular atrophy patient. *Nature* **457**, 277–80 (2009).

[301] Yu, J. *et al.* Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324**, 797–801 (2009).

[302] Bock, C. *et al.* Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–52 (2011).

[303] Mrugala, D. *et al.* Gene expression profile of multipotent mesenchymal stromal cells: Identification of pathways common to TGFbeta3/BMP2-induced chondrogenesis. *Cloning Stem Cells* **11**, 61–76 (2009).

[304] Hamidouche, Z. *et al.* Priming integrin alpha5 promotes human mesenchymal stromal cell osteoblast differentiation and osteogenesis. *Proc Natl Acad Sci U S A* **106**, 18587–91 (2009).

[305] Payton, J. E. *et al.* High throughput digital quantification of mRNA abundance in primary human acute myeloid leukemia samples. *J Clin Invest* **119**, 1714–26 (2009).

[306] Krejci, O. *et al.* p53 signaling in response to increased DNA damage sensitizes AML1-ETO cells to stress-induced death. *Blood* **111**, 2190–9 (2008).

[307] Majeti, R. *et al.* Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc Natl Acad Sci U S A* **106**, 3396–401 (2009).

[308] Dennis, J., G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**, P3 (2003).

[309] Heppner, G. H. Tumor heterogeneity. *Cancer Res* **44**, 2259–65 (1984).

[310] Fidler, I. J. & Hart, I. R. Biological diversity in metastatic neoplasms: origins and implications. *Science* **217**, 998–1003 (1982).

[311] Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883–92 (2012).

[312] Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–45 (2013).

[313] Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501**, 355–64 (2013).

[314] Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–37 (2013).

[315] Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet* **38**, 468–73 (2006).

[316] Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–58 (2013).

[317] Turner, N. C. & Reis-Filho, J. S. Genetic heterogeneity and cancer drug resistance. *Lancet Oncol* **13**, e178–85 (2012).

[318] Sequist, L. V. *et al.* Genotypic and histological evolution of lung cancers acquiring resistance to egfr inhibitors. *Sci Transl Med* **3**, 75ra26 (2011).

[319] de Beca, F. F. *et al.* Cancer stem cells markers CD44, CD24 and ALDH1 in breast cancer special histological types. *J Clin Pathol* **66**, 187–91 (2013).

[320] Bruna, A. *et al.* TGF-beta induces the formation of tumour-initiating cells in claudinlow breast cancer. *Nat Commun* **3**, 1055 (2012).

[321] Taube, J. H. *et al.* Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc Natl Acad Sci U S A* **107**, 15449–54 (2010).

[322] Janssen-Heijnen, M. L. & Coebergh, J. W. The changing epidemiology of lung cancer in europe. *Lung Cancer* **41**, 245–58 (2003).

[323] Hassan, K. A., Chen, G., Kalemkerian, G. P., Wicha, M. S. & Beer, D. G. An embryonic stem cell-like signature identifies poorly differentiated lung adenocarcinoma but not squamous cell carcinoma. *Clin Cancer Res* **15**, 6386–90 (2009).

[324] Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**, 262–72 (2006).

[325] Shedden, K. *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* **14**, 822–7 (2008).

[326] Raponi, M. *et al.* Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* **66**, 7466–72 (2006).

[327] Sato, M. *et al.* Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations. *Mol Cancer Res* **11**, 638–50 (2013).

[328] Goldstraw, P. *et al.* Non-small-cell lung cancer. *Lancet* **378**, 1727–40 (2011).

[329] Skrzypski, M. *et al.* Main histologic types of non-small-cell lung cancer differ in expression of prognosis-related genes. *Clin Lung Cancer* **14**, 666–673 e2 (2013).

[330] Penney, K. L. *et al.* mRNA expression signature of gleason grade predicts lethal prostate cancer. *J Clin Oncol* **29**, 2391–6 (2011).

[331] Weigelt, B. *et al.* Refinement of breast cancer classification by molecular characterization of histological special types. *J Pathol* **216**, 141–50 (2008).

[332] Mizuno, H., Spike, B. T., Wahl, G. M. & Levine, A. J. Inactivation of p53 in breast cancers correlates with stem cell transcriptional signatures. *Proc Natl Acad Sci U S A* **107**, 22745–50 (2010).

[333] Wynder, E. L. & Muscat, J. E. The changing epidemiology of smoking and lung cancer histology. *Environ Health Perspect* **103 Suppl 8**, 143–8 (1995).

[334] Amos, C. I. *et al.* Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* **40**, 616–22 (2008).

[335] Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* **7**, e1002240 (2011).

[336] Miller, L. D. *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* **102**, 13550–5 (2005).

[337] Pawitan, Y. *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res* **7**, R953–64 (2005).

[338] Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clin Cancer Res* **13**, 3207–14 (2007).

[339] Kao, K. J., Chang, K. M., Hsu, H. C. & Huang, A. T. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer* **11**, 143 (2011).

[340] Loi, S. *et al.* Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* **25**, 1239–46 (2007).

[341] Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–9 (2005).

[342] Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–41 (2006).

[343] Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* **68**, 5405–13 (2008).

[344] Viale, G. *et al.* High concordance of protein (by ihc), gene (by fish; her2 only), and microarray readout (by targetprint) of er, pgr, and her2: results from the eortc 10041/big 03-04 mindact trial. *Ann Oncol* **25**, 816–23 (2014).

[345] Yamauchi, M. *et al.* Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage i lung adenocarcinoma. *PLoS One* **7**, e43923 (2012).

[346] Botling, J. *et al.* Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin Cancer Res* **19**, 194–204 (2013).

[347] Der, S. D. *et al.* Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage ia patients. *J Thorac Oncol* **9**, 59–64 (2014).

[348] Kratz, J. R. *et al.* A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. *Lancet* **379**, 823–32 (2012).

[349] Subramanian, J. & Simon, R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst* **102**, 464–74 (2010).

[350] Kleer, C. G. *et al.* EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A* **100**, 11606–11 (2003).

[351] Gonzalez, M. E. *et al.* EZH2 expands breast stem cells through activation of NOTCH1 signaling. *Proc Natl Acad Sci U S A* **111**, 3098–103 (2014).

[352] Takawa, M. *et al.* Validation of the histone methyltransferase EZH2 as a therapeutic target for various types of human cancer and as a prognostic marker. *Cancer Sci* **102**, 1298–305 (2011).

[353] Shao, C. *et al.* Essential role of aldehyde dehydrogenase 1A3 for the maintenance of non-small cell lung cancer stem cells is associated with the STAT3 pathway. *Clin Cancer Res* **20**, 4154–66 (2014).

[354] Takahashi-Yanaga, F. & Kahn, M. Targeting Wnt signaling: can we safely eradicate cancer stem cells? *Clin Cancer Res* **16**, 3153–62 (2010).

[355] Buijs, J. T. *et al.* The BMP2/7 heterodimer inhibits the human breast cancer stem cell subpopulation and bone metastases formation. *Oncogene* **31**, 2164–74 (2012).

[356] Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. *CA Cancer J Clin* **64**, 9–29 (2014).

[357] van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–6 (2002).

[358] Andre, F. & Delaloge, S. First-generation genomic tests for breast cancer treatment. *Lancet Oncol* **11**, 6–7 (2010).

[359] Cardoso, F., Piccart-Gebhart, M., Van't Veer, L. & Rutgers, E. The mindact trial: the first prospective clinical validation of a genomic tool. *Mol Oncol* **1**, 246–51 (2007).

[360] Bergot, E. *et al.* Predictive biomarkers in patients with resected non-small cell lung cancer treated with perioperative chemotherapy. *Eur Respir Rev* **22**, 565–76 (2013).

[361] Xie, Y. & Minna, J. D. A lung cancer molecular prognostic test ready for prime time. *Lancet* **379**, 785–7 (2012).

[362] Bilal, E. *et al.* Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Comput Biol* **9**, e1003047 (2013).

[363] Dowsett, M. *et al.* Comparison of PAM50 risk of recurrence score with oncotype DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J Clin Oncol* **31**, 2783–90 (2013).

[364] Cuzick, J. *et al.* Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the genomic health recurrence score in early breast cancer. *J Clin Oncol* **29**, 4273–8 (2011).

[365] Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160–7 (2009).

[366] Rhodes, D. R. *et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166–80 (2007).

[367] Rustici, G. *et al.* Arrayexpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res* **41**, D987–90 (2013).

[368] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–50 (2005).

[369] Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–40 (2011).

[370] Banerji, C. R., Severini, S., Caldas, C. & Teschendorff, A. E. Intra-tumour signalling entropy determines clinical outcome in breast and lung cancer. *PLoS Comput Biol* **11**, e1004115 (2015).

[371] Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171–8 (2005).

[372] Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488–92 (2005).

[373] Haibe-Kains, B., Desmedt, C., Sotiriou, C. & Bontempi, G. A comparative study of survival models for breast cancer prognostication on microarray data: does a single gene beat them all? *Bioinformatics* **24(19)**, 2200–2208 (2008).

[374] Pierce, G. B. & Speers, W. C. Tumors as caricatures of the process of tissue renewal: prospects for therapy by directing differentiation. *Cancer Res* **48**, 1996–2004 (1988).

[375] Vanderplanck, C. *et al.* The fshd atrophic myotube phenotype is caused by dux4 expression. *PLoS One* **6**, e26820 (2011).

[376] Griggs, R. C. *et al.* Monozygotic twins with facioscapulohumeral dystrophy (fshd): implications for genotype/phenotype correlation. *Muscle Nerve Suppl* S50–5 (1995).

[377] Rahimov, F. *et al.* Transcriptional profiling in facioscapulohumeral muscular dystrophy to identify candidate biomarkers. *Proc Natl Acad Sci U S A* **109**, 16234–9 (2012).

[378] Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–27 (2007).

[379] Tonini, M. M. *et al.* Asymptomatic carriers and gender differences in facioscapulohumeral muscular dystrophy (fshd). *Neuromuscul Disord* **14**, 33–8 (2004).

[380] Knopp, P. *Deciphering the transcriptopnal networks of DUX4: a candidate gene for Facioscapulohumeral muscular dystrophy.* Ph.D. thesis (2011).

[381] Winokur, S. T. *et al.* Facioscapulohumeral muscular dystrophy (fshd) myoblasts demonstrate increased susceptibility to oxidative stress. *Neuromuscul Disord* **13**, 322–33 (2003).

[382] Pau, G., Fuchs, F., Sklyar, O., Boutros, M. & Huber, W. EBImage–an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–81 (2010).

[383] Celegato, B. *et al.* Parallel protein and transcript profiles of fshd patient muscles correlate to the D4Z4 arrangement and reveal a common impairment of slow to fast fibre differentiation and a general deregulation of MyoD-dependent genes. *Proteomics* **6**, 5303–21 (2006).

[384] Banerji, C. R. *et al.* beta-catenin is central to dux4-driven network rewiring in facioscapulohumeral muscular dystrophy. *J R Soc Interface* **12**, 20140797 (2015).

[385] Gordon, M. D. & Nusse, R. Wnt signaling: multiple pathways, multiple receptors, and multiple transcription factors. *J Biol Chem* **281**, 22429–33 (2006).

[386] Kyriakis, J. M. & Avruch, J. Mammalian MAPK signal transduction pathways activated by stress and inflammation: a 10-year update. *Physiol Rev* **92**, 689–737 (2012).

[387] Morgan, M. J. & Liu, Z. G. Reactive oxygen species in TNFalpha-induced signaling and cell death. *Mol Cells* **30**, 1–12 (2010).

[388] Yao, Z. *et al.* Dux4-induced gene expression is the major molecular signature in fshd skeletal muscle. *Hum Mol Genet* (2014).

[389] Block, G. J. *et al.* Wnt/beta-catenin signaling suppresses DUX4 expression and prevents apoptosis of FSHD muscle cells. *Hum Mol Genet* **22**, 4661–72 (2013).

[390] Abu-Baker, A. *et al.* Lithium chloride attenuates cell death in oculopharyngeal muscular dystrophy by perturbing Wnt/beta-catenin pathway. *Cell Death Dis* **4**, e821 (2013).

[391] Kaidi, A., Williams, A. C. & Paraskeva, C. Interaction between beta-catenin and HIF-1 promotes cellular adaptation to hypoxia. *Nat Cell Biol* **9**, 210–7 (2007).

[392] Teng, M. *et al.* Microtubular stability affects pVHL-mediated regulation of HIF-1alpha via the p38/MAPK pathway in hypoxic cardiomyocytes. *PLoS One* **7**, e35017 (2012).

[393] Nakao, N. *et al.* Hydrogen peroxide induces the production of tumor necrosis factor-alpha in RAW 264.7 macrophage cells via activation of p38 and stress-activated protein kinase. *Innate Immun* **14**, 190–6 (2008).

[394] Shi, C. S., Huang, N. N., Harrison, K., Han, S. B. & Kehrl, J. H. The mitogen-activated protein kinase kinase kinase kinase GCKR positively regulates canonical and noncanonical Wnt signaling in B lymphocytes. *Mol Cell Biol* **26**, 6511–21 (2006).

[395] Liu, J. & Lin, A. Role of JNK activation in apoptosis: a double-edged sword. *Cell Res* **15**, 36–42 (2005).

[396] Sun, T. Q. *et al.* PAR-1 is a Dishevelled-associated kinase and a positive regulator of Wnt signalling. *Nat Cell Biol* **3**, 628–36 (2001).

[397] Wang, J. *et al.* A peptide inhibitor of c-Jun N-terminal kinase protects against both aminoglycoside and acoustic trauma-induced auditory hair cell death and hearing loss. *J Neurosci* **23**, 8596–607 (2003).

[398] Zhao, Y. *et al.* The JNK inhibitor D-JNKI-1 blocks apoptotic JNK signaling in brain mitochondria. *Mol Cell Neurosci* **49**, 300–10 (2012).

[399] Bakay, M. *et al.* Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration. *Brain* **129**, 996–1013 (2006).

[400] Zahn, J. M. *et al.* Transcriptional profiling of aging in human muscle reveals a common aging signature. *PLoS Genet* **2**, e115 (2006).

[401] Welle, S., Tawil, R. & Thornton, C. A. Sex-related differences in gene expression in human skeletal muscle. *PLoS One* **3**, e1385 (2008).

[402] Urso, M. L., Scrimgeour, A. G., Chen, Y. W., Thompson, P. D. & Clarkson, P. M. Analysis of human skeletal muscle after 48 h immobilization reveals alterations in mRNA and protein for extracellular matrix components. *J Appl Physiol (1985)* **101**, 1136–48 (2006).

[403] Chen, Y. W. *et al.* Transcriptional pathways associated with skeletal muscle disuse atrophy in humans. *Physiol Genomics* **31**, 510–20 (2007).

[404] Kauffmann, A., Gentleman, R. & Huber, W. arrayqualitymetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–6 (2009).

[405] Hsiao, L. L. *et al.* Correcting for signal saturation errors in the analysis of microarray data. *Biotechniques* **32**, 330–2, 334, 336 (2002).

[406] Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* **6**, e17238 (2011).

[407] Pounds, S. B. Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform* **7**, 25–36 (2006).

[408] Komurov, K., Dursun, S., Erdin, S. & Ram, P. T. NetWalker: a contextual network analysis tool for functional genomics. *BMC Genomics* **13**, 282 (2012).